Josephson-CMOS hybrid memories

Qingguo Liu

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-49
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-49.html

April 25, 2007

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
25 APR 2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Josephson-CMOS hybrid memories	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California at Berkeley,Department of Electrical Engineering and Computer Sciences,Berkeley,CA,94720**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Lack of high-density, fast and memory has been a long-standing problem in su- perconducting digital electronics. Alternative Josephson-junction-based memory cells and peripheral circuits have intrinsic problems which impede the application of that kind of memory. The CMOS-Josephson hybrid memory idea was proposed in 1992 to circumvent this problem. Evaluation of some issues was carried out in the 1990's. In the present work we have designed and demonstrated a 64-kb CMOS-Josephson hybrid memory working at 1 GHz; it has proved to be a promising memory candidate for superconducting high-end computing applications. In order to make the hybrid idea practical, a complete and comprehensive 4 K short-channel CMOS model for digital circuits was developed for this dissertation. It has been used in the design of the memory core and peripheral circuits. A fast hybrid interface circuit that transforms single-?ux-quantum millivolt 2-ps pulse signals to CMOS volt-level signals has been studied and the optimized design has been com- pleted. Using a nonvolatile cryogenic 3-T DRAM cell, the memory core is designed in such a way that the access time is much less than for room-temperature opera- tion, not only because of the performance upgrade due to the temperature decrease but also because of a di?erent reading mechanism. Fabricated by commercial submi- cron CMOS processes (0.25 ?m and 0.18 ?m) and a Nb/AlOx/Nb Josephson-junction process with 2.5 kA/cm2 tunneling current density, the memory chips were bonded together by direct wire-bonding and by ?ip-chip bump-bonding and tested at low frequencies, high frequencies, respectively. A subnanosecond access time is obtained both from simulations and experiments. The power for an interface circuit is mea- sured to be 0.6 mW. The memory core consumes even less power. The total power consumption depends on the operation mode of the memory and is calculated to be 10 mW for reading and 28 mW for writing. Simulations also indicate that, with more advanced semiconductor and superconductor technologies, larger and faster hybrid memories are expected to be achievable in the future.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**198**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Josephson-CMOS Hybrid Memories

by

Qingguo Liu

B.S. (Nanjing University) 1999
M.S. (Nanjing University) 2002

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Electrical Engineering
and Computer Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Theodore Van Duzer, Chair
Professor Ahmad Bahai
Professor John Clarke

Spring 2007

The dissertation of Qingguo Liu is approved:

Chair Date

Date

Date

University of California, Berkeley

Spring 2007

Josephson-CMOS Hybrid Memories

Abstract

Josephson-CMOS Hybrid Memories

by

Qingguo Liu

Doctor of Philosophy in Engineering-Electrical Engineering

and Computer Sciences

University of California, Berkeley

Professor Theodore Van Duzer, Chair

Lack of high-density, fast and memory has been a long-standing problem in superconducting digital electronics. Alternative Josephson-junction-based memory cells and peripheral circuits have intrinsic problems which impede the application of that kind of memory. The CMOS-Josephson hybrid memory idea was proposed in 1992 to circumvent this problem. Evaluation of some issues was carried out in the 1990's. In the present work we have designed and demonstrated a 64-kb CMOS-Josephson hybrid memory working at 1 GHz; it has proved to be a promising memory candidate for superconducting high-end computing applications.

In order to make the hybrid idea practical, a complete and comprehensive 4 K short-channel CMOS model for digital circuits was developed for this dissertation. It

has been used in the design of the memory core and peripheral circuits. A fast hybrid interface circuit that transforms single-flux-quantum millivolt 2-ps pulse signals to CMOS volt-level signals has been studied and the optimized design has been completed. Using a nonvolatile cryogenic 3-T DRAM cell, the memory core is designed in such a way that the access time is much less than for room-temperature operation, not only because of the performance upgrade due to the temperature decrease, but also because of a different reading mechanism. Fabricated by commercial submicron CMOS processes (0.25 μm and 0.18 μm) and a Nb/AlO_x/Nb Josephson-junction process with 2.5 kA/cm^2 tunneling current density, the memory chips were bonded together by direct wire-bonding and by flip-chip bump-bonding and tested at low frequencies, high frequencies, respectively. A subnanosecond access time is obtained both from simulations and experiments. The power for an interface circuit is measured to be 0.6 mW. The memory core consumes even less power. The total power consumption depends on the operation mode of the memory and is calculated to be 10 mW for reading and 28 mW for writing. Simulations also indicate that, with more advanced semiconductor and superconductor technologies, larger and faster hybrid memories are expected to be achievable in the future.

Professor Theodore Van Duzer
Dissertation Committee Chair

To my parents, Tongyin liu and Xinqiu Xu

to my wife and my son, Chengjuan Sun and Theodore C. Liu

Acknowledgments

In the past five years, I spent all my time making good *memories*.

This five years in Berkeley are going to be the most memorable years in my life. I have been very fortunate to be surrounded by outstanding and inspiring professors and encouraging friends and colleagues who have continuously offered me supports and encouragement. I sincerely appreciate their contributions to my professional and personal growth.

First and foremost, I would like to thank my advisor Professor Theodore Van Duzer, for giving me the opportunity to study and do research at Berkeley, for his support and guidance in my research, for his advisory and help in my personal matters. I have been benefited not only from his wisdom and knowledge, but also his unique and creative ways of analyze and troubleshoot practical problems, which will become invaluable treasures in my life. He is always a decent and considerate gentleman, providing adequate help and shedding lights in front of my way.

Along with Professor Van Duzer, I would like to thank Professor Ahamd Bahai, Professor John Clarke for being my dissertation committee. I also thank Professor Vivek Subramanian, Professor John S. Smith, Professor John Clarke, and Professor Theodore Van Duzer for serving as my qualify examination committee members. And I sincerely appreciate Professor and chief technologist of National Semiconductor Corporation (NSC), Ahmad Bahai, for giving me a great chance to make a smooth transition from academy to industry, and to work with outstanding people in NSlabs.

I would like to thank Dr. Xiaofan Meng, for everything he has done for me. I could hardly express my appreciation in a few words. His excellency in device physics and process helps me a lot. And his experience and lab techniques are one of the most valuable treasures of our group. I would like to thank Prof. Nobuyuki Yoshikawa of Yokohama National University, Japan for the inspiring discussions and the fruitful cooperations. Also, I want to thank Dr. Kan Fujiwara for his dedication in testing experiments, especially for preparing the flip-chip bump-bonding chip sets. And I would like to thank Dr. Steven Whiteley for his CAD support and fruitful discussions. I want to express my appreciation to my collegues in NSlabs, especially to Wei Ma and Ali Djabbari, for providing me a friendly working environment and a flexible schedule.

This work was supported by National Semiconductor Corporation and Office of Navy Research (ONR). The chip fabrication was performed by National Semiconductor Corporation and Superconducting Research Laboratory of NEC (NEC-SRL). I would like to thank Dr. Steven Michael Lanzisera of Berkeley, Dr. Jianhui Zhang and Dr. Jim Wieser of NSC for their help on CMOS layout. And I would like to thank Dr. Hidaka of NEC-SRL for his generous support on JJ chip fabrication and bump-bonding preparation.

Finally, I would like to thank my parents and my wife. Their love and support are the most important asset in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Tales of two technologies

1.1 Overview of CMOS and Josephson technologies

1.1.1 Overview of CMOS technology

Ever since the invention of first semiconductor transistor by William Shockley, John Bardeen and Walter Brattain in 1947 at Bell Laboratories, and especially after the first CMOS circuit was invented in 1963 by Frank Wanlass at Fairchild Semiconductor [1], semiconductors have increasingly taken over the electronics world. Originally a low-power but slow alternative to TTL, CMOS found early adopters in the watch industry and in other fields where battery life was more important than speed. Some twenty-five years later, CMOS has become the predominant technology in digital integrated circuits. This is essentially because area occupation, operating speed, energy efficiency, and manufacturing costs have benefited and continue to benefit from the geometric downsizing that comes with every new generation of semiconductor manufacturing processes, well-known as Moore's Law [2]. In addition, the simplicity and comparatively low power dissipation of CMOS circuits have allowed for integration densities not possible on the basis of bipolar junction transistors.

Besides the digital world, CMOS plays an important role in the analog world. Despite of the lower cut-off frequency, CMOS finds a position in RF circuits due to the much lower cost than that of any bipolar technologies. As of 2006, most advanced CMOS digital products are manufactured by using a 65 nm process with a power

supply of 1.3 V, and a 45 nm process with a less than 1 V power supply will be soon sent to the product line [3]; the most advanced analog products are manufactured by using a 130 nm process with operating frequencies up to 60 GHz [4].

1.1.2 Overview of Josephson junction technology

Josephson technology, on the other hand, shares almost the same length of history as CMOS technology, but has had much less impact on the electronics industry, largely because of the need of refrigeration.

1.1.2.1 Josephson junction and fabrications

The active device in superconducting electronics is a junction between two superconductors with a weak link, which is weak enough to allow interference of the electron pair wave functions. Predicted by B. D. Josephson in 1962 [5], under this condition, electron pairs can pass through the junction without having any applied voltage, causing a supercurrent.

It can be derived from quantum mechanics that,

$$I = I_c \sin \phi, \tag{1.1}$$

$$V_J = \frac{\hbar}{2e} \frac{\partial \phi}{\partial t} \tag{1.2}$$

where ϕ is the phase difference of the wave functions across the junction, I_c is the maximum current passing through the junction without voltage drop, called the

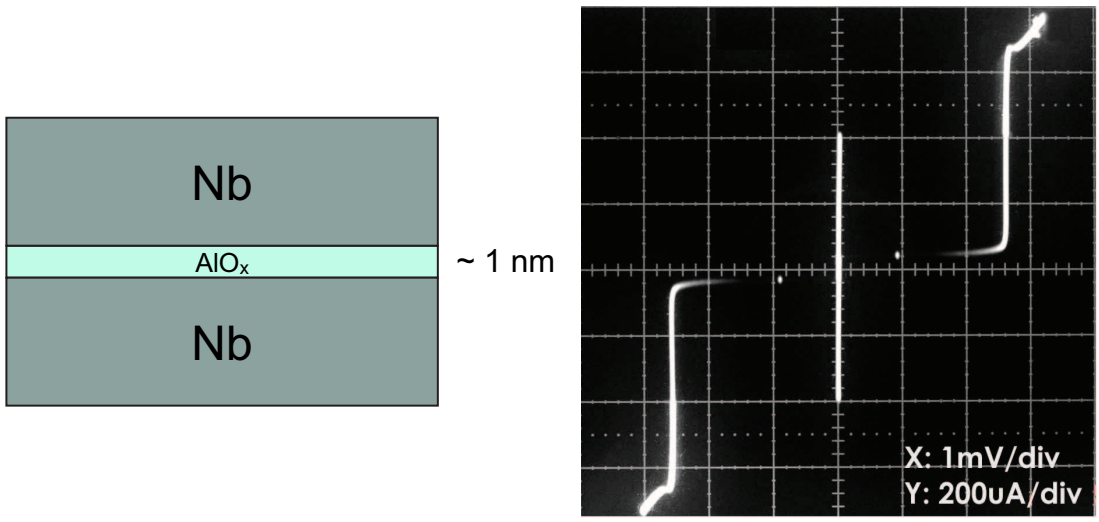


Figure 1.1: The structure of a SIS Josephson junction and its typical I-V characteristic.

criticalcurrent. In Eq. 1.2, the $2e$ term represents the electron pairs in superconductors. These two equations were first derived by B. D. Josephson and are called the Josephson equations.

There are numerous ways to implement such a weak link between two superconductors, including metal or semiconductor links, grain boundaries, very narrow constrictions, damaged regions, and, most prominently, insulating tunnel barriers. Junctions made by superconductor-insulator-superconductor (SIS) sandwich structure, as shown in Fig. 1.1 (a), are the workhorses in superconducting electronics because of the highly developed state and robustness. The most popular junction technology is the Nb/AlO_x/Nb technology. Two niobium films deposited on silicon wafers are weak-linked by an aluminum oxide layer with a typical thickness of 1-2

nm, and the critical current density depends heavily on the thickness of the junction.

A typical I-V curve of Nb junction is shown in Fig. 1.1 (b). When the bias current is less than the critical current, there is no voltage drop across the junction; when the bias current goes beyond the critical current, the junction switches to the so-called voltage state with a gap voltage $V_g = 2.6$ mV drop across it. After that, if we decrease the current, there is a hysteresis in the I-V curve due to the large capacitance between two electrodes. If the junction is shunted with a resistor, the hysteresis will be smaller. The smaller resistance value, the smaller the hysteresis.

After the first Nb/AlO_x/Nb process was invented and the RSFQ logic family was introduced, however, the situation changed dramatically. The Nb/AlO_x/Nb junction was developed by Gurvitch *et al.* [16]. The fabrication process was further improved by Morohashi *et al.*. And X. Meng *et al.* [7] implemented a new approach called light anodization such that submicron junctions with small variations are possible. The introduction of Nb/AlO_x/Nb junctions resolved most of the problems in process reliability. And the high quality and excellent uniformity of the junction characteristics enabled high-speed operation with large margins.

1.1.2.2 Switching of a single SIS junction

As predicted by the Josephson equations, the current in a junction cannot be larger than the critical current, otherwise the junction will switch to the voltage state. A typical Josephson junction switching process is approximately an RLC charging

process. In this simple model, a nonlinear inductor, a physical junction capacitor, and a junction resistor are connected in parallel, and a current source is connected to them to supply current to switch the junction. After the current goes beyond the critical current of the junction, the voltage across the junction tries to rise to the gap voltage (V_g); however, the presence of the capacitor prohibits the voltage increasing instantaneously because the capacitor requires some charging current, which comes from the bias current minus the junction and resistor currents. If the charging time is much longer than the LC oscillation cycle time, which is the case in typical Nb processes, the charging time can be simply written as,

$$t_J = \frac{V_g C_J}{I_c} \tag{1.3}$$

The I_c in the denominator represents the average charging current: because of the oscillatory current in the junction, the current that charges the capacitor can vary at the Josephson frequency between twice the critical current and zero. So the whole equation is simply the time to charge a capacitor up to V_g with a constant current source I_c.

Fig. 1.2 shows the simulation results both with WRSPICE and HSPICE that verify the analysis above. In Fig. 1.2, τ_{on} represents the time until the phase of the junction increases one more π to get close to $3\pi/2$, and τ_{rise} is the charging time. In a 2.5 kA/cm^2 Nb process, the charging time is 6.16 ps based on Eq. 1.3 and 6.20 ps based on the simulation. We can see the LC oscillation on the curve, which confirms

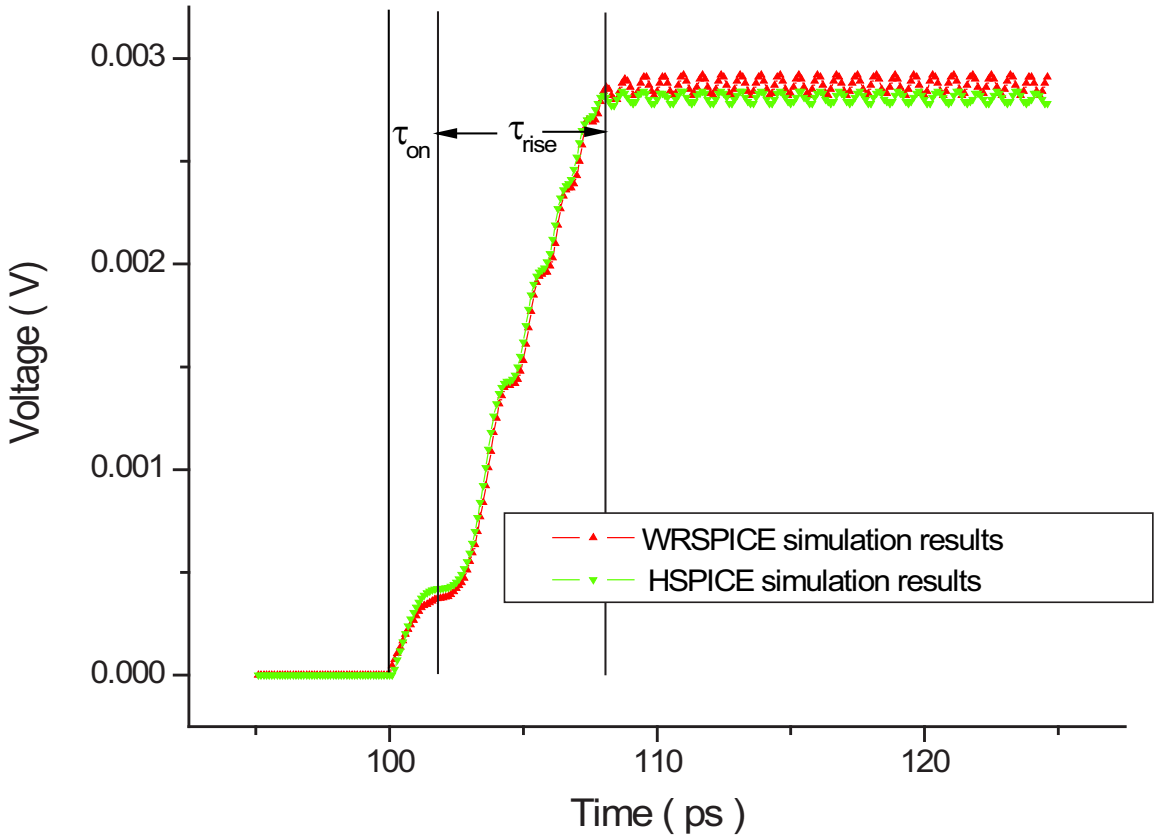


Figure 1.2: The WRSPICE and HSPICE simulation results for a single-junction switching. The RC charging time (τ_{rise}) is about four times larger than the oscillation time. And the results of the two simulators show a good consistency.

that in this process, the period is much smaller than the switching time. The figure also shows the consistency between WRSPICE and HSPICE simulations.

This charging process can be viewed from the perspective of the pendulum analog [49]. In the pendulum analog, the angular velocity of the pendulum represents the voltage and the angle off vertical indicates the current in the junction as a fraction of I_c according to the Josephson equation Eq. 1.1.

When the angle of the pendulum is increased beyond 90 degrees by some applied torque that represents the applied current in the Josephson junction, the pendulum starts to roll. Due to the moment of inertia (which represents the junction capacitor), the angular velocity (voltage) cannot increase to the gap voltage instantaneously; the time that it takes depends on the total energy on the pendulum after it reaches continuous rotation and how fast the applied torque can pump energy into it, in other words, the power. The time can be written as :

$$t_J = \frac{E}{P} = \frac{I_p \omega^2}{\tau_{applied} \omega} \Rightarrow \frac{C_J V_g}{I_c} \tag{1.4}$$

where I_p is the moment of inertia of the pendulum, ω is its maximum phase velocity of the pendulum, and $\tau_{applied}$ is the applied torque. This equation agrees with Eq. 1.3.

1.1.2.3 Circuit applications of Josephson junctions

Josephson junctions can be used in many circuit applications. Between the prediction and discovery of the Josephson effect [5] and the implementation of the first Nb

process [6] [7], Josephson technology was only employed for analog applications. The most successful applications include, but are not limited to, SQUID magnetometers [10], SIS mixers [11] [12], and the Josephson voltage standard [8] [9]. There were no digital applications. There are three primary reasons for that. The first one is the integration level. Each device on the list above is quite simple and requires only one or a few Josephson junctions, except for the voltage standard, which requires only a single long series array of uniform junctions. For a simple digital circuit, however, the integration level is much higher if it is to serve a useful application. The complexity puts great demands on fabrication variations. Given the primitive state of superconducting integrated circuit fabrication facilities and technologies, it was difficult to make a large complex digital circuit.

The second reason is the logic style. Before the rapid-single-flux-quantum (RSFQ) logic family was introduced [13] [14], the so-called voltage-state logic style ruled the Josephson digital world. This logic style, shares the same basic idea as CMOS logic, using voltage to represent a digital "1" or a "0". It was faster than the CMOS logic at that time. However, the problems with this logic are intrinsic. The main problems are the "punchthrough" (malfunction due to random failure to reset to zero-voltage state [15]), which limits the operating speed of a voltage-state logic circuit, and the junction nonlinearity, which decreases the operation margins. The third one is that it was difficult to make a high-speed cache memory at that time. As a result, IBM, after about 15 years, stopped their big Josephson computer project in 1983, causing

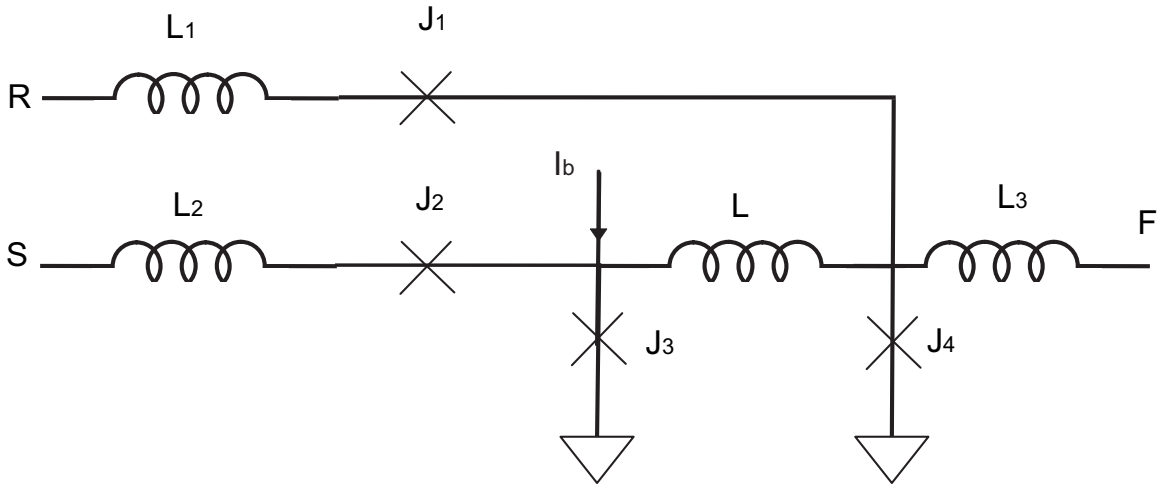


Figure 1.3: A simple RSFQ RS flip-flop. [13]

a general slowing of the research in Josephson-junction-based digital electronics.

The RSFQ (currently meaning Rapid Single Flux Quantum) logic family was invented by K. Likharev *et al.* [13] and this logic family employs a new concept of digitizing, which enables us to utilize a quantum properties of Josephson circuits and, therefore, make faster circuits. The RSFQ signals are voltage pulses with the integral over time being a magnetic flux quantum ($\Phi_0 = 2.07 \times 10^{-15} Wb$). And a typical pulse will have 0.5 - 1 mV height and several ps duration.

Fig. 1.3 shows a simple RSFQ RS flip-flop. The J_3, L J_4 loop is a two-junction interferometer with $I_c L = 1.25\Phi_0$ so it can store a flux quantum. Initially, the bias current almost all flows to J_3 and it looks like there is a counterclockwise circulating current in the loop superimposed on equal currents in J_3 and J_4. When a pulse arrives on the S input, its current passes through J_2 and causes J_3 to switch and

the circulating current is transferred to J_4, leaving only a small current in J_3. The circulating current in the loop now is clockwise, representing a stored "1". Then when a pulse arrives at R, its current causes J_4 to switch and resets the loop current to counterclockwise, representing a stored "0". All the junctions in RSFQ logic are shunted with small resistors, therefore, the RSFQ circuits are mostly combinations of junctions, resistors and inductors.

Both voltage-state logic and RSFQ logic can have great speed performance. Many different logic circuits were fabricated and operated at much higher speeds than semiconductor circuits. The Josephson voltage-state gate performed a record of 1.5-ps gate delay at a power dissipation of only 12 pW [18]. The world's first Josephson microprocessor operated at a clock frequency of 770 MHz [19]. As of 2006, the fastest small Josephson RSFQ microprocessor worked at 15 GHz [20]. And a 15-bit ADC works at 20 GHz sampling frequency has been designed and demonstrated [21].

1.2 Problems with CMOS technologies

After the CMOS is scaled into the short-channel region, i.e., the channel length is less than 1 μm, short-channel effects come into play and bring many challenges and benefits. The benefits are obvious. First of all, shrinking of the channel length and width leads to less area consumption and, therefore, lower costs. The shrinking of area also reduces the parasitic capacitance and, therefore, the intrinsic delay. Secondly, the scaling includes thinning the gate oxide. The supply voltage has to be reduced

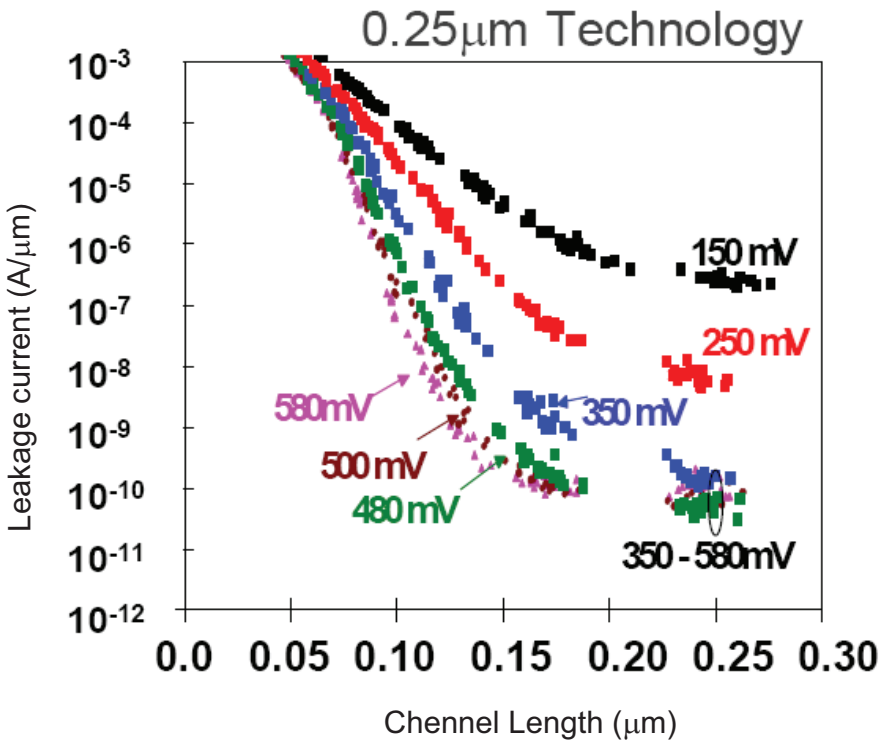


Figure 1.4: Off-state leakage current versus channel length for 0.25 μm transistors with different threshold voltages. [23]

accordingly because of the higher electrical field in the silicon if the voltage remains unchanged. The voltage reduction helps a lot in power consumption. With continuing scaling, more problems are coming as well. In order to solve the problems, new structures and/or new materials have been proposed. The paper "The End of the CMOS scaling" by Skotnicki, T. *et al*. [22] published in 2005 gives a good review on this topic. The conclusion is that the scaling can continue this way for some time, but it becomes more and more difficult and requires more and more research attention. In that paper, there are still some unresolved problems.

The most important problem faced by advanced CMOS is power dissipation, espe-

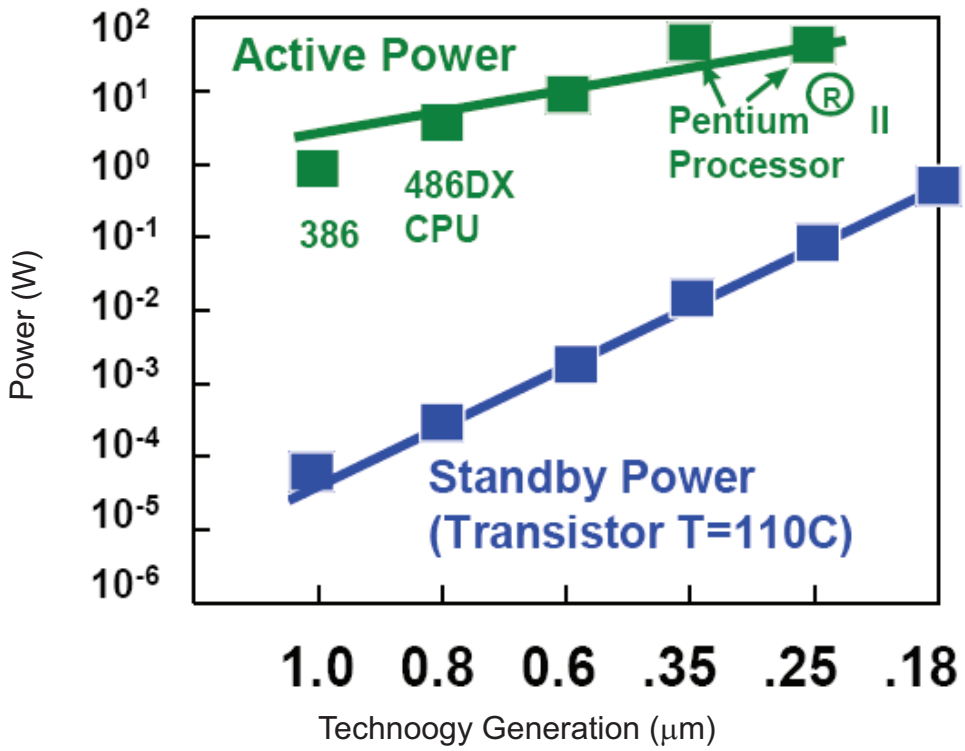


Figure 1.5: Active and standby power trends for Intels technologies. [23]

cially for deeper scaling. The power consumption for a digital CMOS circuit has two parts. One is the static power, which is supposed to be extremely small as one of the main benefits of CMOS technology; the other is the dynamic power, which is CV^2f, where C is the capacitance and V is the power supply voltage and f is the operating frequency. When the channel length is longer than 1 μm, the static power is indeed very small compared with the dynamic power. However, when the technology gets more and more advanced, the static power becomes increasingly important.

The static power mainly comes from subthreshold leakage current. Subthreshold leakage is intrinsic in silicon MOSFET operation and is related to the device thresh-

old voltage. Subthreshold off-state leakage versus channel length characteristics are shown in Fig. 1.4 [23]. When the supply voltage and threshold voltage is scaled down, the subthreshold current (off current) gets larger; even with reduced V_{DD}, the static power increases. The active and standby power trends for Intel's process technologies are shown in Fig. 1.5 [23]. In this figure, we can find that the standby power for 1 μm technology was 0.01% of active power, but approaches 10% of active power in 0.1 μm technology. In the Pentium 4 processor made by Intel, the static current causes more than 20% of the total power dissipation, even after clever circuit design in order to minimize the leakage current.

In order to limit the increase of standby power, threshold voltages need to increase. However, this increase strongly affects device performance because of reduced gate over-drive. To maintain acceptable leakage values, the V_T of transistors will need to increase by more than 0.25 V. The best subthreshold swing one can make is about 60 mV/dec at room temperature, and for a 0.18 μm process, it is normal that the swing goes up to 80 - 100 mV/dec. For a threshold voltage of 0.25 V, it means the turn-off leakage current in a transistor is only two or three orders smaller than the turn-on current when gate voltage is right above threshold voltage. The threshold voltage cannot be any smaller or the static leakage current will be a disaster.

In order to maintain the circuit speed, the power supply has to be larger than four times of the threshold voltage. And since the threshold voltage is limited by the static power issue, the scaling of the voltage is slowing down, causing a dynamic

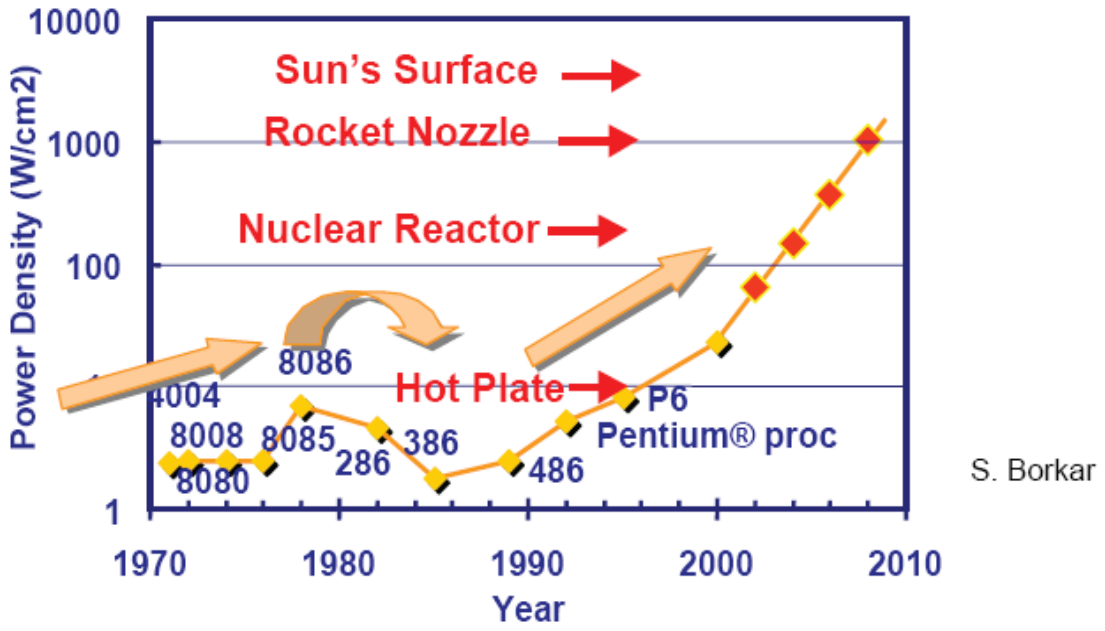


Figure 1.6: Normal power density trend for CMOS digital circuits. Courtesy of Professor Borivoje Nikonic, UC Berkeley [32].

power increase. Taking all this into account, the power consumption is a big problem for more and more advanced CMOS technology. Since the area decreases, the power density is even a more important problem facing by CMOS scaling.

Fig. 1.6 shows the power density trend of CMOS digital circuits [32]. If we keep the trend, we will have to use more powerful cooling systems with a high cost in the next five years capable of cooling the hot chips. Even worse, in the next ten years, we may not be able to find a cooling system in order to cool down the super-hot chips. Besides, with the scaling of supply voltage and increasing current, the power management is becoming more and more challenging for very large scale digital circuits like microprocessors.

With technology scaling, interconnections in CMOS circuits, both on-chip and

off-chip, play an increasingly important role. There are two parameters associated with an interconnection: resistance and capacitance. For an on-chip wire which is made of copper/aluminum alloy, the resistance is in the order of 100 mΩ per square and the capacitance is about 0.2 fF/μm^2. For a long wire, the capacitance is more important than the resistance because the resistance is much less than the equivalent resistance of the driving NMOS or PMOS, while the capacitance is much larger than the parasitic capacitance of the NMOS and PMOS. In order to deal with a long wire, inverter buffers are introduced midway to minimize the delay. When the technology keeps being scaled down, the buffers deliver more and more current, and the voltage drop in the resistance gets larger and larger, compared with the smaller and smaller supply voltage. And the Ldi/dt noise becomes a more series problem. Although it may not be a big problem in the next couple of years, the interconnection eventually will be a limiting factor for CMOS technology.

1.3 Problems with Josephson technologies

1.3.1 Refrigeration

Refrigeration has been the most important impediment to the application of superconductive electronics, including Josephson-junction-based technology. Despite efforts people have made in order to find superconductors with higher critical temperature, there is no superconductor that works at room temperature. Even if one day

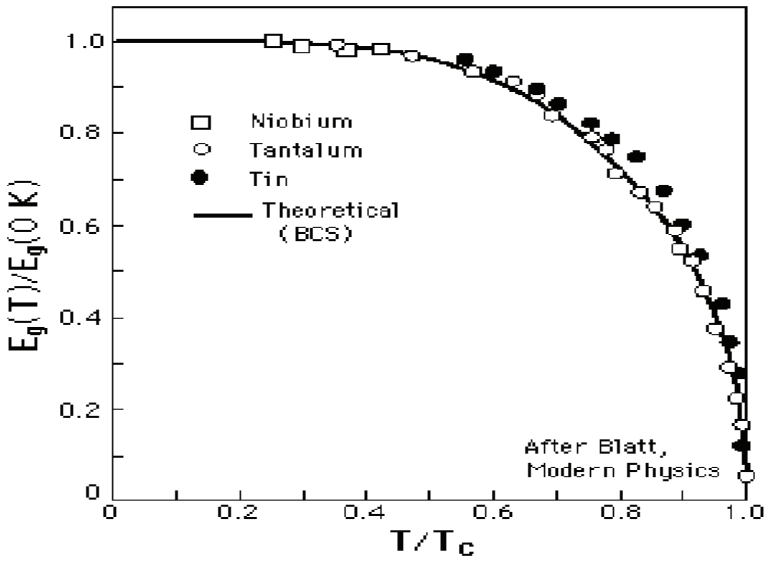


Figure 1.7: The energy gap versus temperature curve for typical superconductors. When the temperature is lower than $(1/2)T_c$, the energy gap is close to the 0 K value and is not very sensitive to temperature.

a room-temperature superconductor ($T_c \geq 300\ K$) should be discovered, the working condition would have to be half of the critical temperature. The typical energy gap versus temperature curve shows the reasons, as shown in Fig. 1.7. First of all, when temperature is higher than half of the critical temperature, the energy gap drops, causing lowered characteristic voltage. (The characteristic voltage is a measure of the density of the superconductor pairs and, therefore, of all special superconductor phenomena.) More importantly, the derivative of the gap energy with temperature is higher, meaning that a small environmental temperature change may lead to a large change in characteristic voltage, causing an unstable operation of circuits. Therefore, in order to have a superconductor that would work well at room temperature, the superconductor would require a critical temperature of at least 600 K.

There are two basic types of refrigerators used for electronics systems working below 150 K [24]. One type compresses a gas at room temperature, decreasing its entropy with little change in the enthalpy, and makes it to do work at low temperature, thus providing the desired refrigeration. The other type, so called Joule-Thomson refrigerator, also compresses a gas at room temperature, but the gas is imperfect and the enthalpy is decreased. The gas is allowed to expand irreversibly without doing any work at room temperature. The gas cools then providing the desired refrigeration. Most of the popular refrigerators like Sterling-cycle and the Gifford-McMahon-cycle refrigerators are in the category of type one. And the Joule-Thomson type coolers are typically used in smaller system with less than 1 W cooling capacity. The coefficient of performance (COP) of a refrigerator is defined as the ratio of the cooling capacity to the input power. The ideal COP is the Carnot COP where a Carnot recycle is being performed:

$$\eta = \frac{T_r}{T_a - T_r}, \tag{1.5}$$

where T_r is the refrigerator temperature and T_a is ambient temperature. However, real refrigerators can only achieve a fraction of the Carnot COP. The typical 4 K refrigerators only have 5% COP of an ideal Carnot refrigerator. And typical 77 K refrigerators can increase the number up to 25 %. That implies that for every milliwatt of cooling capacity provided at 4 K, 1.5 watt of input power is required for the refrigeration; and every milliwatt of cooling power provided at 77 K requires

about 16 milliwatt of input power for refrigeration. Or, in other word, in order to beat other competitors in terms of wall power, circuits based on Josephson technology must require at least three orders less power than the competitors. In addition to the power loss in refrigeration, the size and weight of refrigerators is a problem, especially to portable consumer electronics. So far, no successful superconductor applications are for portable consumer electronics.

Although the power spent on refrigeration increases the total power of a superconducting system by a factor of 1000, for superconducting circuits based on voltage-state logic or RSFQ logic, the power consumed by the Josephson junctions is still small. An RSFQ circuit, for example, consumes energy of $I_c\Phi_0$ when pumping in or out a quantum flux. Taking a critical current of 200 μA, operating frequency of 100 GHz, the power consumption would be 40 nanowatts, multiplying the factor of 1000, the input power is only 40 μW. However, that is only the power consumed by junction switching. The junctions have to be biased and the bias current source will consume some power. It turns out that the static bias current power is much larger than the switching power. Some design tricks can be played in order to have a less power hungry bias current source; however, those tricks sacrifice delay time, the operation margin or both, and they only work for Josephson-CMOS hybrid circuits.

1.3.2 Integration level

An important impediment to the application of superconducting digital circuits is the low integration level, especially compared with the highly-matured very large scale CMOS circuits. As discussed in previous paragraphs, the integration level of superconducting circuits was extremely poor before the emergence of the Nb process. Even after the Nb process was developed to a more advanced level, the integration level of superconducting circuits is still much lower than that of the semiconductor competitor. As of 2006, one single most advanced CMOS chip has more than 10^9 gates on it, while the most advanced superconducting chip only has about 10^4 gates.

Large-scale circuits require high consistency within chips, and, for acceptable yield, chip-to-chip, wafer-to-wafer, and run-to-run consistency is needed. For CMOS circuits, the most important parameters are the threshold voltage and the physical gate size. The lithography and doping are the most important reasons for the parameter variations. For superconducting circuits, the most important parameter is the critical current, which is basically determined by the size of the junction and the thickness of the barrier. In a typical Nb process, the barrier is an approximately 1-nm-thick aluminum oxide made by thermal oxidation of the deposited aluminum. In other words, the barrier only contains about several layers of atoms. Due to the exponential relationship between barrier thickness and the critical current, any small change in the barrier thickness will cause a lot more change in the value of the critical current. That is one of the main reasons for the low integration level. Besides thickness, the

quality of the oxide is very important. Any defects or traps in the aluminum oxide barrier layer will change the barrier height and, thus, the tunneling properties.

Although there are problems that limit the consistency of superconducting chips, after years of efforts, the critical current spread has been improved substantially. For a standard, well-developed 2.5 kA/cm^2 Nb process, there have been reports that the critical variation is less than 6 $\sigma = 5$ %. This is a very good number and even better than the numbers for some of the most advanced CMOS processes. However, this does not mean the integration level of superconducting circuits should be at least as high as CMOS circuits. The reason is circuit robustness. Having the same or better process spread, superconducting circuits still cannot achieve the same integration level as CMOS circuits, because CMOS circuits have much larger margins than superconducting circuits. In other words, CMOS circuits, especially digital ones, are much more robust that superconducting circuits.

As mentioned in Sec. 1.2, the voltage-state circuits have low margins due to the switching dynamics and the nonlinearity of Josephson junctions. The punchthrough effect, for example, is a contributor to this problem. For RSFQ circuits, there are still limited margins, even though the junction shunt resistance suppresses the punchthrough effect a great deal. The main reason for the small margins is the poor fain-in and fan-out capability, especially compared to CMOS logic. For CMOS logic, the fan-out could be any number, the only price is time to drive the following capacitance and, therefore, the large delay associated with that. The margins of the

multi-fan-out CMOS circuit are still the same as the margins of the circuit that has fan-out of one. RSFQ circuits, however, have different situation. The more fan-out a circuit has, the poorer margins the circuit will have.

The lack of three-terminal devices is one of the key reasons for the fan-out problem. Unlike a metal-oxide-semiconductor field-effect transistor (MOSFET), a Josephson junction has only two terminals, the current flowing through the two terminals acts as the control element, similar to the gate voltage in a MOSFET. The difference is, for a semiconductor MOSFET, the gate capacitor isolates the interaction between the control element and the other two terminals, at least at low frequency. While for a Josephson junction, there is no such an isolation mechanism. Once a junction is connected with other junctions or passive devices, the bias condition changes to a different level, leading to lower bias margins. The other important reason is the quantum property of the flux quantum. The output of an SFQ pulse cannot be split into half and transmitted into the next gates. A splitter circuit must be inserted in order to provide fanout. And this splitter circuit has lower margin for more fan-out requirement.

1.3.3 Memory bottleneck

Memories are essential for computation and extensive efforts have been devoted to make suitable memories for Josephson-junction-based computation systems. Random Access Memories (RAMs) are especially critical for high-end computation. Un-

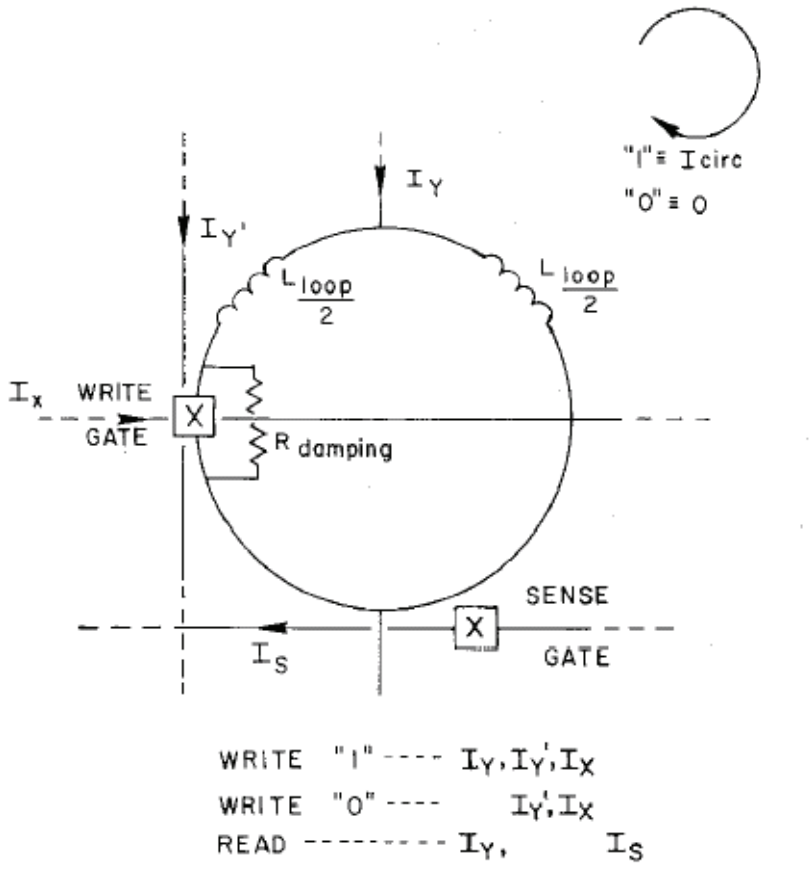


Figure 1.8: Henkels's nondestructive read-out (NDRO) memory cell. After [25].

fortunately, lack of a large, fast and sufficiently robust Josephson-junction-based RAM with high bit yield has been a long-standing problem since the first Josephson junction-based memory cell was developed by IBM in 1970's [25] and was the main reason for discontinuing their project in 1983.

In the following paragraphs, three important types of Josephson RAM cells will be discussed. These are the non-destructive read-out cell developed by H. Henkels at

IBM [25], the variable-threshold cell developed by I. Kurosawa at ETL [26], and the vortex-transitional cell developed by S. Tahara at NEC [27].

As shown in Fig. 1.8, Henkel's cell uses an interferometer for each of the write and read operations, in addition to the storage loop, where information is stored as magnetic flux. The two states of the cell are zero current, representing a "0", and a clockwise current, representing a "1". A "1" is written by a coincidence of a data current I_d and control currents I_x and I_y, and removing the data current after the removals of the control current. A "0" is written by the application of control current I_x and I_y, and dissipating the circulating current, if any exists. The stored data is read by a coincidence of a data current and a sense current I_s. If a "1" is stored, the sense current will be large enough to switch the read gate. If a "0" is stored, the sense current is not enough to switch the gate. The product of the loop inductance and the critical current must be larger than $(n+1/2)\Phi_0$ and less than $(n+1)\Phi_0$, where n is an integer, preferably zero. If n is larger than 1, the number of stored flux quanta may vary, causing a margin decrease of the sense gate. The main problems for this memory cell are the half-selection problem and the large area. Half selection arises when the column decoder and row decoder send x and y current to select a certain cell, the other cells in the same row are half selected by x current and cells in the same column are half selected by y current.

Kurosawa's cell, sometimes called a variable-threshold memory cell, is a much smaller single-flux-quantum memory cell, and is shown in Fig. 1.9. The cell consists

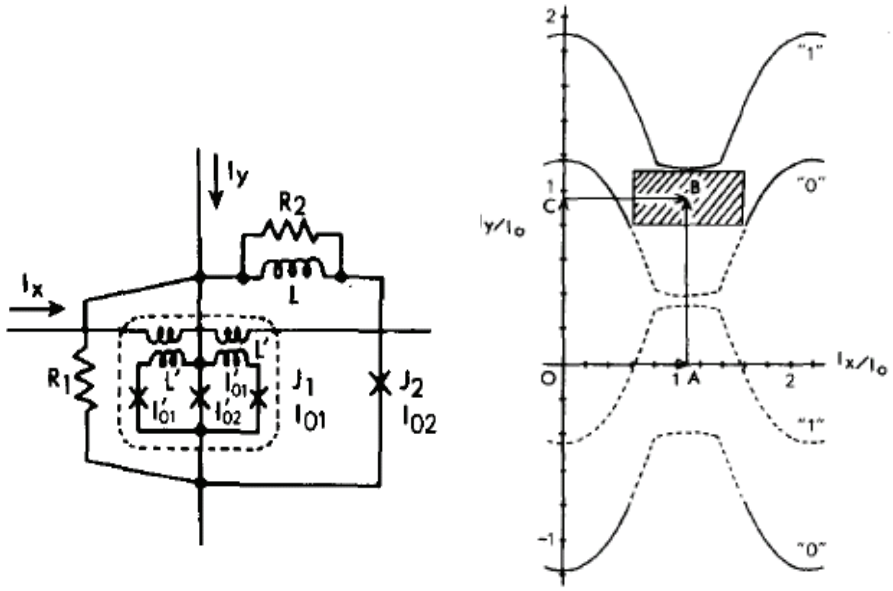


Figure 1.9: ETL's variable-threshold memory cell. The solid line represents the vortex-to-voltage transition region, while the dashed lines represent vortex-to-vortex port. R_2 is the damping resistor. After [26].

of a write gate J_1 composed of a three junction SQUID, a single junction J_2 and the storage loop formed by the inductance L. The word current I_x is applied to the cell as the control current for J_1 gate, and controls the maximum critical current of J_1. The critical current of J_2 is chosen to be the maximum critical current when $I_x = 0$. The loop inductance is designed such that the loop can barely hold a single flux quantum. The cell changes the threshold curve depending on the presence or absence of a single flux quantum in the cell. The stored information is defined as a "1" if there is a flux quantum, otherwise a "0".

For writing a "1" into a cell, I_x current is applied; if the cell stores a "1" before the writing, nothing happens; if the cell stored a "0", the stored flux gets lost through J_1, resulting an O-A path on the threshold curve. For writing a "0", I_y is applied after I_x, following a O-A-B path on the curve. This operation opens J_1 causing I_y to flow to J_2, so no flux is trapped in the loop. All transitions in the writing process are vortex-to-vortex processes, and no voltage state is involved. For the reading process, however, I_y is applied before I_x, following O-C-B path on the threshold curve. If there is a "1" stored, nothing happens. And if there is a "1" stored, path C-B crosses the threshold curve, causing the cell to switch to the voltage state. Thus, the readout process is destructive and a rewriting procedure is necessary. This cell is a little smaller than the Henkel's cell based on same Nb process. However, because the column and row currents are still there, there is still the half-selection problem. Its destructive reading process makes this cell less attractive.

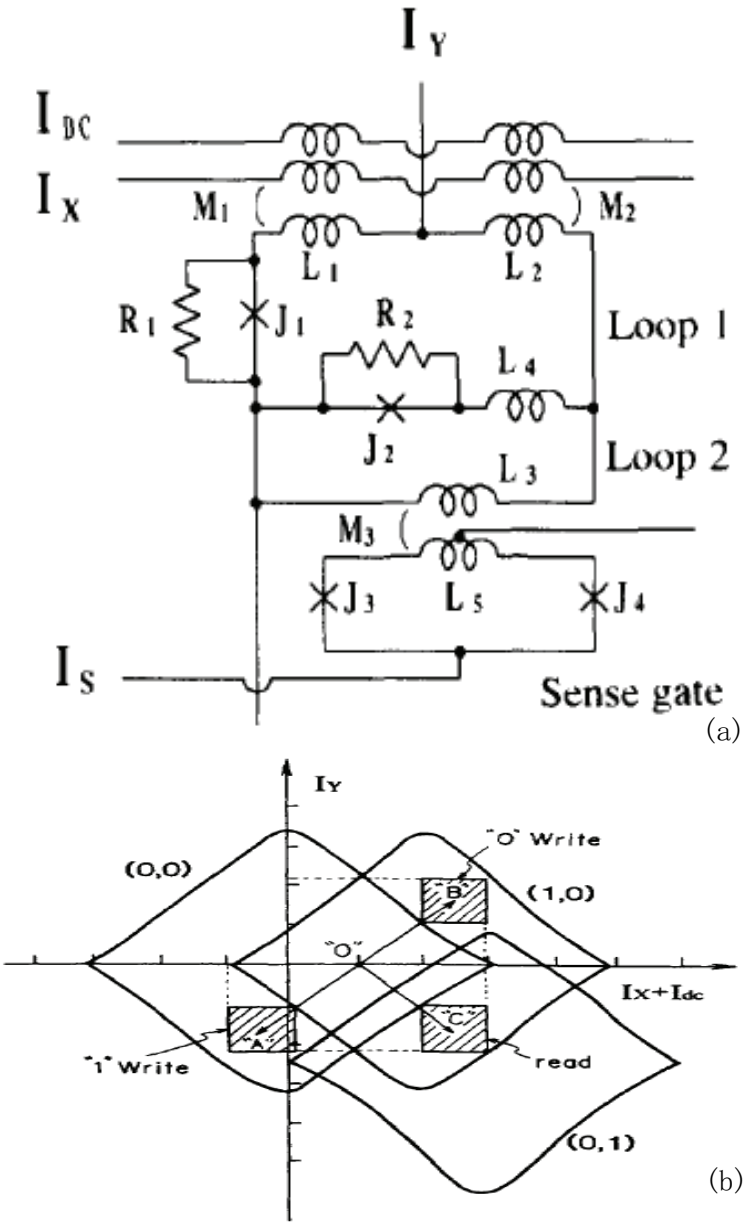


Figure 1.10: The vortex transitional memory cell demonstrated by NEC. (a) equivalent circuit and (b) the threshold curve. The mode (m,n) means that m flux quantum are stored in the data-storage loop and n flux quantum are stored in the read-enable loop. After [28]

The vortex-transitional memory cell shown in Fig. 1.10 performs nondestructive read-out operations without compromising the operating margins. The cell consists of a data-storage loop, a read-enable loop, and a sense gate. When both I_x and I_y are positive, the junction J_1 causes a flux quantum to enter the data-storage loop, corresponding to a "0" in this cell; when both I_x and I_y are negative, there are no flux quanta in the data storage loop, meaning a "1" in this cell. When positive I_x, negative I_y, and a sense current are applied, the cell performs a reading operation. If a "1" is stored, the junction J_2 undergoes a vortex-to-vortex transition, and a flux quantum is injected into the read-enable loop. So the datum can be read out by the switching of the sense gate. The reported cell size was 55 μm $\times$ 55 μm in a 3 μm process. And the measured margins were $\pm$23%. The important drawback of this memory cell array is the half-selection of memory cells.

Density is one of the main issues of Josephson-junction-based memories. All the junction-based memory cells that have been demonstrated successfully so far consist not only Josephson junctions, but also inductors and resistors. These passive elements occupy a lot of area, thus limiting the density of the memory. It is normal that the area for these passive elements dominates, so when the technology is scaled, the area of memory cells does not shrink very much. For SQUID-based memory cell, the situation gets even worse. When the junction size is scaled down, the junction capacitance shrinks, so the shunt resistance has to be larger in order to make the McCumber number less than unity [29]. Therefore, the presence of inductors and

resistors limits the density of memories as well as the scalability. Typically the size of a junction-based memory cell is about 50 μm $\times$ 50 μm [27] for a standard 2.5 kA/cm^2 Nb/AlO_x/Nb process. This size is huge compared with the size of a DRAM cell [30] or even an SRAM cell [31] in a commercial 0.25 μm CMOS technology. And with further technology scaling, the cell size of a semiconductor memory can be even smaller. The area difference between a Josphson memory cell and a semiconductor memory cell is getting larger and larger. The other problem for Josephson-junction-based memories is the small margins compared to semiconductor competitors, especially for larger sizes. The yield of a memory depends on its robustness of the memory cell and peripheral circuits. CMOS logic is famous for the robustness. Furthermore, large-scale superconducting digital circuits have low margins for two main reasons. One is that the technology spread is not good enough for a large-scale system and the other is the problematic intrinsic switching dynamic of any individual circuit. As of 2006, the largest working memory was the NEC's 4-kb memory. [44]

1.4 Thesis Overview

The CMOS and Josephson memories both have some intrinsic problems. In general, CMOS circuits have great robustness and scalability and have already dominated the digital world, but the power density will be a big problem in the future. Josephson circuits, on the other hand, have great performance with less power consumed. But the low margin and low integration level prohibit Josephson technology from being

widely used in large-scale applications, especially for high-end computing applications. For different applications, there are different aspects which determine which technology is preferable. However, memory application for a high-end superconductor-based computation system, requires high density, robustness, and good performance. One likely solution is the Josephson-CMOS hybrid memory, in which the CMOS memory is operated at 4 K in order to be intimately connected to the Josephson processing circuits. This approach utilizes the robustness and high density of the CMOS technology. The signal-level differences are accommodated by the input amplification and superconductor bit line sensors. Most importantly, it is so far the only likely memory solution for a high-end computation system.

In this dissertation, the design and demonstration of the hybrid memory will be addressed as follows. Chapter 2 presents the low-temperature semiconductor physics and CMOS characteristics, and a 4 K BSIM 3 model is established in order to simulate the hybrid circuits. Designs and simulations are discussed in Chapter 3. Every circuit in the hybrid memory system is studied by simulation and final optimization is discussed. All measurements are presented in Chapter 4; both low-frequency functionality tests and high-frequency delay measurements are shown. Chapter 5 discusses future hybrid memories and concludes this dissertation.

Chapter 2

Short-Channel MOSFETs at cryogenic temperatures

2.1 Introduction

The hybrid memory is the most promising solution for the superconducting memory problem. The semiconductor-superconductor hybrid, or more specifically, the CMOS-Josephson hybrid circuit solution, is based on the performance of two technologies. In order to achieve a compact hybrid, which means to put two types of circuits together to minimize the interconnection degradation, the CMOS circuits must be cooled to the temperature at which the superconducting circuits can work well, which is 4 K for Nb Josephson circuits. There then arises the following questions about 4 K CMOS: Does 4 K CMOS work at all? Are we able to develop new models to successfully represent CMOS at extremely low temperature? Fortunately, answers to these questions are yes, and we can use low-temperature CMOS to build our hybrid memory. And all these questions will be answered in detail in this chapter.

Low-temperature CMOS is not a new research topic. There are many applications that require CMOS circuits to be operated at low-temperature. And even CMOS design designers and device engineers have done research on low-temperature CMOS for the purpose of performance upgrade. However, due to the convenience of using liquid nitrogen, compared with liquid helium, most low-temperature CMOS research was done with liquid nitrogen at 77 K. Only a small fraction of low-temperature CMOS research was based on liquid helium, which has a boiling temperature of 4.2 K. Besides, for most previous low-temperature CMOS research, the focus was on long-channel CMOS devices and circuits, because of the process limitation at that

time.

In this chapter, important parameters of MOSFETs at low temperatures will be studied both theoretically and experimentally. For experiments, the samples were fabricated using the CMOS-8 0.25 μm and CMOS-9 0.18 μm (twin well) processes provided by National Semiconductor Corporation (NSC). The minimum size of a transistor using CMOS-8 0.25 μm process is 0.25 μm $\times$ 0.3 μm. For our experiments, both NMOS and PMOS devices with short channel lengths (0.25 μm) and long channel lengths (10 μm) were fabricated on a test chip. Static I-V characterization was carried out on these discrete devices at room temperature and at cryogenic temperatures (77 K and 4.2 K), and the dynamic properties are studied as well. A comprehensive understanding of 4 K short-channel CMOS will be addressed, both from the device-physics point of view and the circuit-design point of view, and we established a complete 4 K BSIM3 model for a commercial CMOS process and verified it by experiments.

2.2 Steady-state carrier properties at 4 K

In conventional silicon MOSFETs, n-channel and p-channel devices are doped with boron and phosphorus, respectively. Both phosphorus donor levels and boron acceptor levels are located about 45 meV from the corresponding band edge. At room temperature, thermal energy is large enough to excite electrons or holes into the conduction band or the valence band, in a short enough time, leaving behind ionized

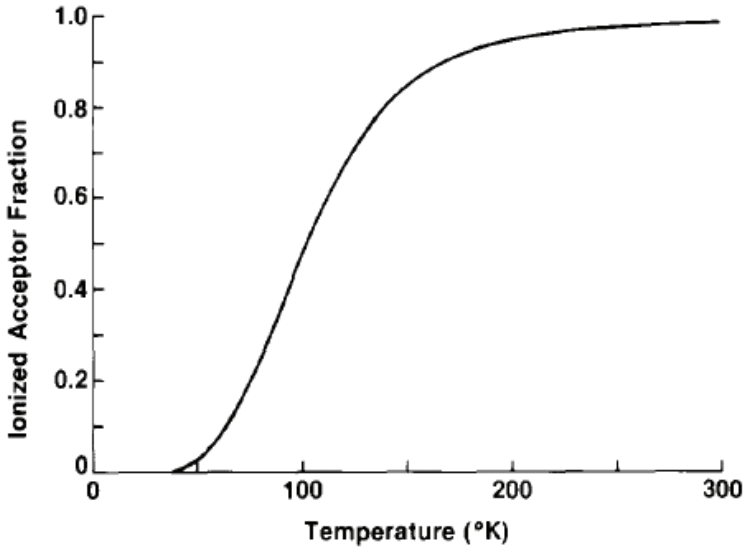


Figure 2.1: Ionized impurity fraction versus temperature in nondegenerately boron-doped silicon. [33]

impurities. The carrier concentration as a function of temperature is easily determined by computing the charge neutrality condition in doped silicon. A calculation of the ionized acceptor concentration in non-degenerate boron doped silicon ($2.8 \times 10^{16} cm^{-3}$) is illustrated in Fig. 2.1 [33]. As the temperature drops below 100 K, the ionized impurities act as shallow traps; due to the decrease of thermal excitation energy, carriers begin to occupy these shallow levels. As the temperature keeps decreasing, there are fewer and fewer carriers staying in the bands until at 0 K, when there are no carriers remaining in the bands; this is known as carrier freeze-out. According to the temperature, the freeze-out can be categorized into two regions, strong freeze-out and weak freeze-out. Weak freeze-out happens between about 30 K to 100 K,

when the thermal energy is somewhat weak, after some short time (often called dwell time), carriers can pick up enough thermal energy to escape from the trap. Note that this is a dynamic process, meaning there are some carriers being trapped back to the donor/acceptor levels at the same time. Strong freeze-out, on the other hand, happens when temperature is below 30 K. At such a low temperature, the thermal energy is very low and carriers have to wait a long dwell time in order to pick up enough energy to escape the trap, which makes the ionization practically impossible. Theoretical calculation indicates an exponential relationship between dwell time and temperature. [34] For example, the dwell time for 25 K is as large as one second, and the dwell time for 10 K is about 10^{16} seconds!

However, for devices at extremely low temperature, hard freeze-out is not immutable, the situation can be changed by an applied electrical field. In a frozen-out MOSFET, the application of a field of sufficient strength can induce ionization of trapped carriers and depletion-region formation. Two types of field-dependent ionization are possible. One mechanism is Poole-Frenkel ionization [35]. Consider a carrier trapped in a shallow level. With no field applied, the trapped carrier is in a funnel shaped potential with a certain ionization energy. When a field is applied, the potential well tilts and the barrier to thermal ionization is lowered, which makes it easier for the carriers to escape the potential well. The change in the magnitude of the barrier depends entirely on the field and can be calculated. [35] This mechanism does not provide energy to carriers, but only helps to lower the potential and there-

fore to lower the dwell time to a small value corresponding to the applied field. The other possible field-dependent ionization mechanism is tunneling. As in the case of field-assisted thermal ionization, an applied field alters the funnel-shaped potential presented to the trapped carrier; at some sufficient field strength, the carrier is able to tunnel through the barrier into the band and be swept away by the field. Tunneling is a weakly temperature-dependent process, and can thus be labeled as field-induced ionization.

In a 0.25 μm CMOS process, the minimum channel length of a MOSFET is about 0.2 μm and the supply voltage is 2.5 V, which gives a field of 1.25 MV/m. In this field, the shape of the potential well is tilted greatly and the tunneling effect is dominant. Since the tunneling process is a weakly temperature-dependent process, the ionization could be considered as temperature independent. That explains why there are still depletion regions under gates of MOSFET's, even at 4.2 K.

With sufficiently high doping, no freeze-out occurs. When the doping level is so high that the energy levels are degenerate, the carriers do not need to pick up thermal energy to escape from the impurity ions. This is often called a Mott transition, where a semiconductor turns into a metal-like conductor and all the carriers are totally free like electrons in metals. An intuitive explanation for this effect is that the impurity level is so high that the effective electron radii are increased enough to interfere to each other, resulting in lower required energy to escape from the potential. The doping level where the Mott transition [36] happens varies with doping materials as

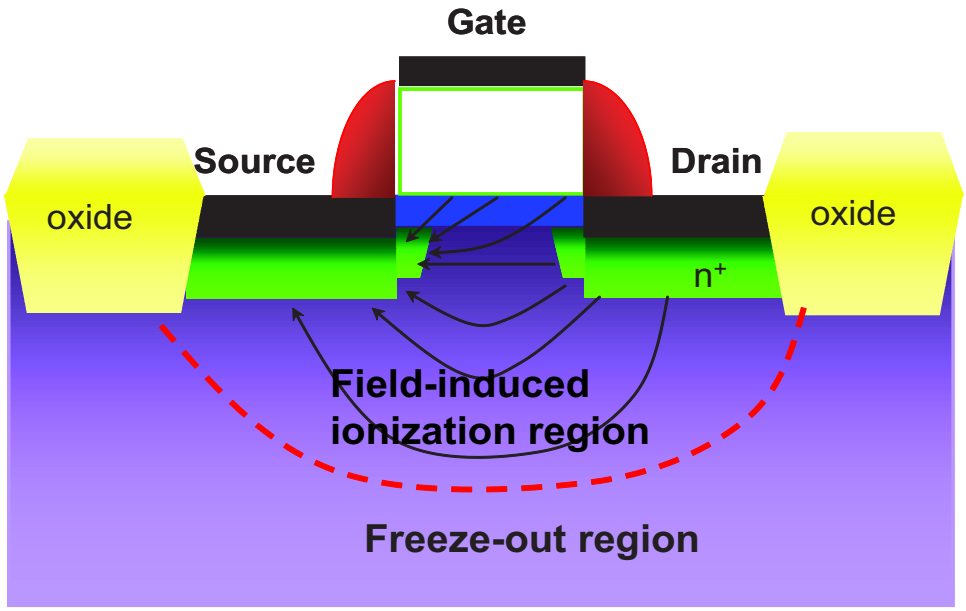


Figure 2.2: Electric fields in an NMOS transistor with gate and drain voltages applied. Only a small region under the gate is frozen-out-free region due to the field induced ionization.

well as with semiconductor materials. For silicon doped with boron and phosphorus, this transition doping level is about $10^{18} cm^{-3}$. Source and drain regions in a modern CMOS process are doped higher than that; therefore, the low temperature has little effect on those regions.

Fig. 2.2 shows a cross section view of a traditional submicron MOSFET with gate and drain voltages applied. Most of the substrate is frozen-out and only the substrate under the gate benefits from the field-dependent ionization. The frozen substrate forms a perfect isolation between different on-chip circuits and effectively prevents circuits from suffering from the latch-up effect (the inadvertent creation of a low-impedance path between the power supply rails that causes latching of, for example, an inverter into the always-conducting state), which bothers most 300 K

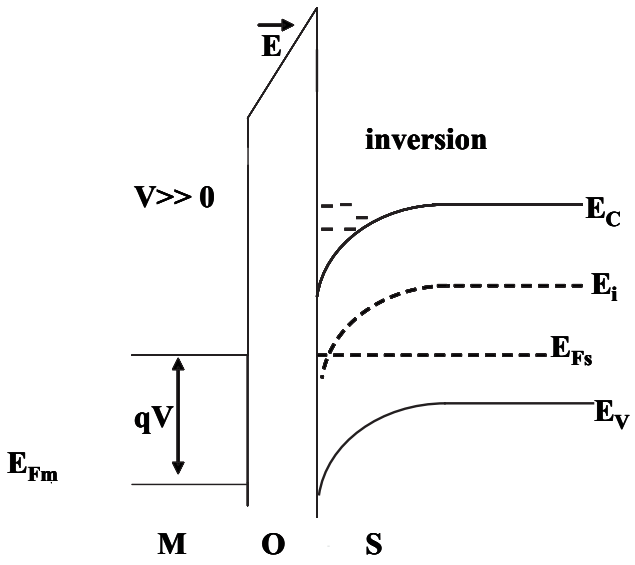


Figure 2.3: Band diagram of a n-type MOSFET biased at threshold voltage. The substrate semiconductor at the interface "looks" like the same n-type as the bulk substrate semiconductor "looks" like p-type.

circuits with submicron devices and requires more precaution in design.

2.3 Threshold Voltage

As one of the most important parameters of a MOSFET, its threshold voltage is defined as the lowest gate voltage at which an inversion region forms in the substrate (body) of the transistor. In an NMOS device, the substrate of the transistor is composed of p-type silicon which has more holes than electrons. When a voltage is applied on the gate, the electric field causes electrons in the substrate to become concentrated at the surface region between the gate oxide and the substrate. Threshold voltage is the gate voltage at which the concentration of surface electrons becomes

equal to that of the holes in the bulk; an inversion region is created as the voltage increases. In terms of energy bands, inversion happens when the surface fermi level (E_f)is higher than the intrinsic level (E_i), and the potential difference is the same as the difference between the intrinsic level and the Fermi level in the bulk p-substrate, as shown in Fig. 2.3.

Based on the definition, the threshold voltage of an n-type MOSFET can be written as

$$V_T = V_{FB} + 2\varphi_s + \frac{\sqrt{2\epsilon_s q N_a (2\varphi_s + V_{SB})}}{C_{ox}}, \tag{2.1}$$

where V_{FB} represents the flat-band voltage, which is the work-function difference between gate material and the substrate; the $2\varphi_s$ term represents the voltage difference that switches the Fermi level and the intrinsic level; and the third term indicates how much depletion charge is on a gate oxide capacitor, where V_{SB} is the source to bulk voltage. In these three terms, the flat-band voltage is a weak function of temperature, and the other two terms are strong functions of temperature due to the relationship between surface potential and temperature.

$$\varphi_s = \frac{kT}{q} \ln(\frac{N_a}{n_i}) \tag{2.2}$$

$$n_i = \sqrt{N_c N_v} e^{-\frac{Eg}{2kT}} \tag{2.3}$$

In Eq. 2.2, note that although it is not absolutely right to assume that all dopant

impurities are ionized at any given temperature, it is still valid for the case where there is a gate voltage applied to help the ionization, even at 4.2 K. Consider the temperature coefficient of the surface potential,

$$\frac{d\varphi_s}{dT} = \frac{kT}{q}\frac{n_i}{N_a}\frac{dn_i}{dT} + \frac{k}{q}ln(\frac{N_a}{n_i}) \tag{2.4}$$

Because n_i is a stronger function of temperature, the kT/q term does not dominate in $\frac{d\phi_s}{dT}$. Thus the surface potential has a negative temperature coefficient, which is the main part of the threshold-voltage temperature coefficient. In the band diagram, the Fermi level of a p-type semiconductor goes down when the temperature goes down, representing an increasing surface potential. When the temperature decreases to extremely low values, the Fermi level merges with the valence band and stops there [33]. The surface potential at that temperature is then $E_g/2$. Therefore the 4 K threshold voltage of a n-type MOSFET can be written as

$$V_T = V_{FB} + E_g + \frac{\sqrt{2\epsilon_s q N_a(2\varphi_s + V_{SB})}}{C_{ox}} \approx \frac{\sqrt{2\epsilon_s q N_a(2\varphi_s + V_{SB})}}{C_{ox}} \tag{2.5}$$

The first two terms cancel out if an n^+-poly silicon is used as the gate material, which is a typical situation in modern CMOS process.

The threshold voltage of a p-type MOSFET is

$$V_T = V_{FB} - E_g - \frac{\sqrt{2\epsilon_s q N_d(2\varphi_s + V_{BS})}}{C_{ox}} \tag{2.6}$$

and in a typical modern process, p^+-poly is used as the gate material, ending up with PMOS and NMOS threshold voltages, symmetrical about zero.

Due to the increase of the surface potential at low temperature, an increase in NMOS threshold voltage is expected. And at 4 K, it is expected that the threshold voltage is proportional to the square root of the doping level. Due to the symmetry, the threshold voltage of a PMOS is decreased.

In experiments to measure the threshold voltage, it is defined differently for long-channel devices and short-channel devices. And because of the square law of MOSFET current, the threshold voltage of a long-channel device is often extrapolated from an I-V curve with $V_{DS} = V_{GS}$. However, due to short-channel-effects, this method is not valid in short-channel devices. Instead, people use the constant-value definition (the gate voltage at which the drain-source current is equal to 300 nA/μm). In this chapter, traditional square-law extrapolation method is used to extract the threshold voltage of long-channel devices and the constant-value definition is used to define threshold voltage for short-channel devices.

Fig. 2.4 shows the measured threshold voltages for two different devices at different temperatures. Both long-channel and short-channel n-type MOSFETs suffer from threshold-voltage increase at low temperatures. And threshold voltages of short-channel devices are smaller than those of long-channel devices. This is well known as the short-channel effect, and it can be explained by a simple model proposed first by Yau [37]. For short-channel devices, because the channel length becomes comparable to the depletion width, some modifications need to be made to compensate this geometrical effect. The basic idea follows. The qualitative analysis considers the

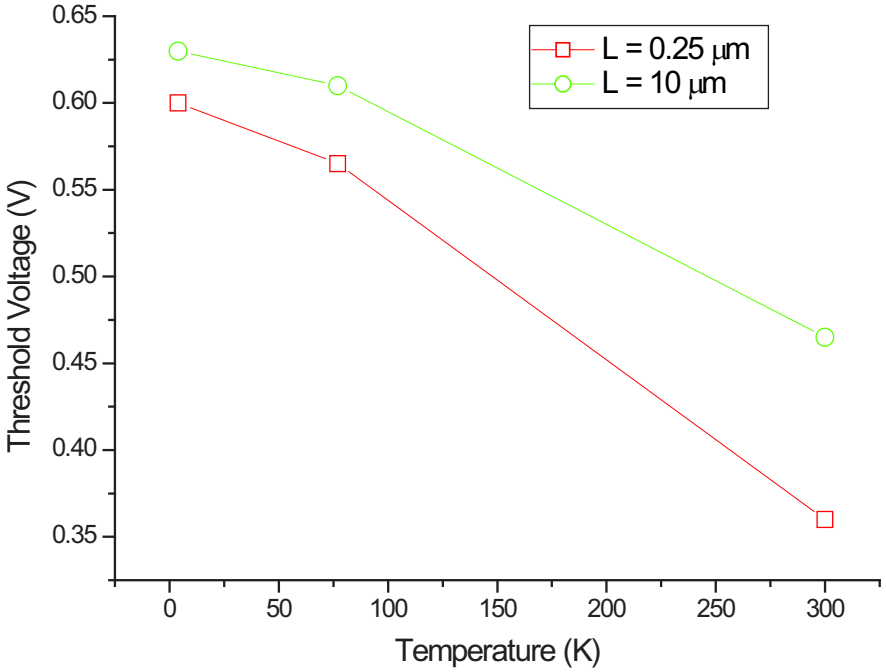


Figure 2.4: Threshold voltages of N-type MOSFETs at different temperatures.

depletion charges distributed evenly inside a cube with the size W, L, W_d, where W and L are the width and length of the gate and W_d is the depletion width. This model is just too simple. Yau proposed that the depletion charge area is not an exact box underneath the gate; instead, it is a box without upper corners. Therefore, there will be a threshold reduction written as

$$\Delta V_T = \frac{qN_A W_d X_j}{C_{ox} L}\left(\sqrt{\frac{2W_d}{X_j} + 1} - 1\right), \tag{2.7}$$

where X_j is the junction depth.

Threshold voltage variation is critical to very-large-scale integrated circuits (VLSI), especially for low-voltage operations. As the technology scaling continues, it is more

and more difficult to maintain small threshold-voltage variation. Due to the difficulty of making a low-temperature automatic probe station, there are no statistical results for 4 K threshold-voltage-variations. However, about 20 devices with same design size have been measured at room temperature as well as at 4 K manually. Although the sample number is too few to make a statistical argument, we still can conclude tentatively that the 4 K variation is at least not worse than that at room temperature.

2.4 Mobility and velocity saturation

Carrier mobility is very important in device performance. It determines how fast an electron or a hole can move for a given electric field. Consider a carrier is moving in a forrest of lattice atoms, and an electric field is applied to accelerate it. According to Newton's law, the carrier will speed up until the field is turned off. However, due to the lattice environment, the carrier will suffer loss of momentum. After several collision and acceleration processes, the carrier velocity remains stable and the momentum gained from the electric field gets totally lost due to the scattering. At high temperatures, the relaxation time is approximately equal to the collision time, while at low temperature, the relaxation time can be many times longer than the collision time. Therefore,

$$Eq\tau = m^* v \tag{2.8}$$

$$\mu = \frac{v}{E} = \frac{q\tau}{m^*}, \tag{2.9}$$

where m^* is the effective mass of the carrier. From the above equations, the mobility of a carrier in a semiconductor depends on the effective mass and the time between two scattering events, the collision time. The effective mass depends on the periodic potential wells, which is basically the lattice structure; and the collision time depends on not only the lattice structure, but also other parameters, such as impurity doping level, defects, and surface roughness, etc.

When temperature changes, the mobility of the carriers changes. It is easy to understand intuitively that decreasing temperature lowers the lattice vibrations and, therefore, decreases the possibility of a collision event (phonon scattering), which is one of the main reasons of the mobility increasing. Fig. 2.5 shows the measured mobility versus applied gate voltage, at different temperatures. At room temperature, the phonon scattering and the surface roughness scattering are the two main mechanisms limiting the carrier transport, as shown in the room temperature curve in Fig. 2.5. As gate voltage increases, electrons get more involved with the surface scattering; therefore, the mobility drops. At low temperatures, especially at cryogenic temperatures, the thermal vibrations of lattice are so small that phonon scattering is not a dominant limiting factor. Instead, Coulomb scattering becomes more important. That is the reason at low temperature the mobility curve increases dramatically with the gate voltage when it is small. When the gate voltage is large enough to attract

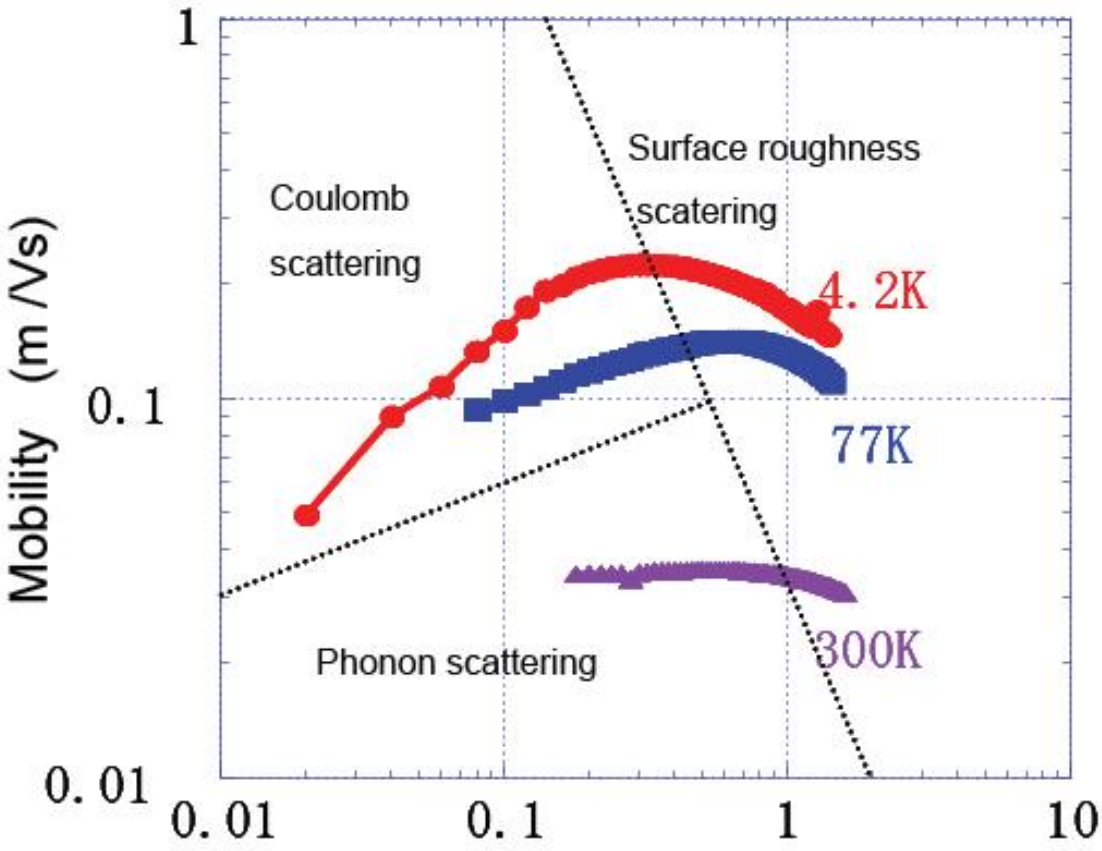


Figure 2.5: The measured low-field mobility of electrons in NMOS devices at different temperatures.

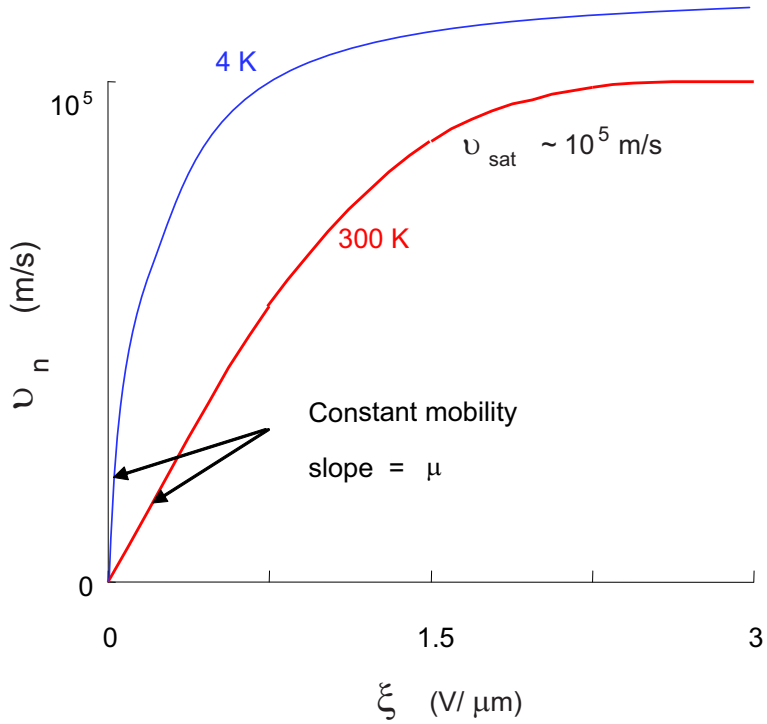


Figure 2.6: Electron velocity versus applied field at room temperature and at 4 K, based on the empirical model. The slope of the low field curve is the low-field mobility. Sharper slope at 4 K means much higher mobility at 4 K.

more and more electrons to the surface, the surface roughness scattering kicks in and play an important role to limit the mobility. On average, the mobility at 4.2 K is more than 10 times higher that that of room temperature, which suggests advantageous high-speed operations at 4 K.

The mobility measured in last section is the carrier drift characteristic at low electric field. At higher field, the drift velocity saturates. And the velocity can be simply modeled in an empirical model

$$v = \frac{\mu_{eff} E}{1 + (E/E_{sat})}, \quad when\ E \leq E_{sat} \tag{2.10}$$

$$v = v_s, \quad when\ E \geq E_{sat} \tag{2.11}$$

As the technology is advancing, the applied field along the channel is getting larger and larger. For a modern CMOS process, the velocity saturation happens earlier than in less advanced processes. It is the velocity saturation that limits the speed performance of a transistor, rather than the low-field mobility. Unlike the mobility, the saturation velocity increases only by about 50% from 300 K to 4 K, and limits the drain current, as shown in Fig. 2.6. The velocity saturation occurs when the field is so high that the optical phonon interaction is dominant. The saturated velocity can be written as

$$v_{sat} \propto \sqrt{\frac{E_p}{m*}}, \tag{2.12}$$

where E_p is the optical phonon energy and m* is the effective mass of electrons. The optical phonon energy increases when temperature decreases, and it increases slowly. That can explain that the improvement of velocity saturation at 4 K is not as phenomenal as the improvement of the low-field mobility.

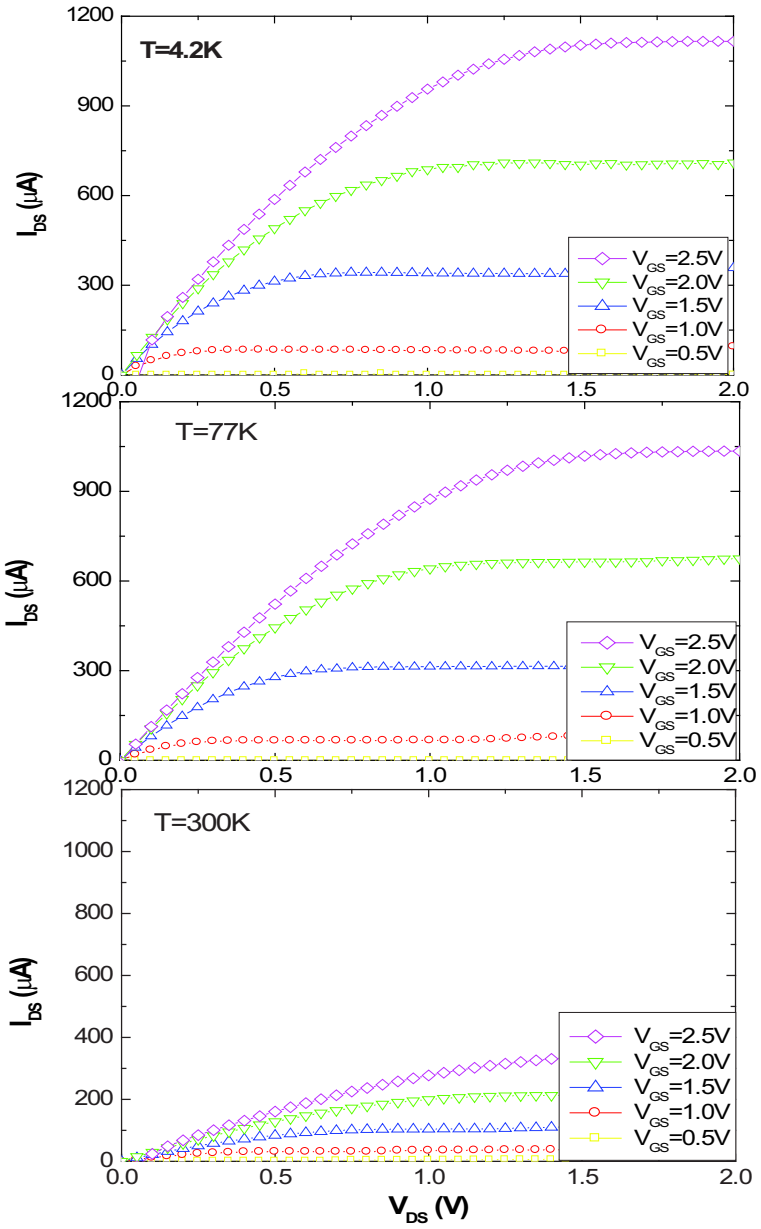


Figure 2.7: Measured I-V curves of a long-channel NMOS device at different temperatures. The saturation current at 4 K increases by a factor of three, compared with the one at room temperature due to the facts that the mobility increases dramatically and the velocity saturation occurs late.

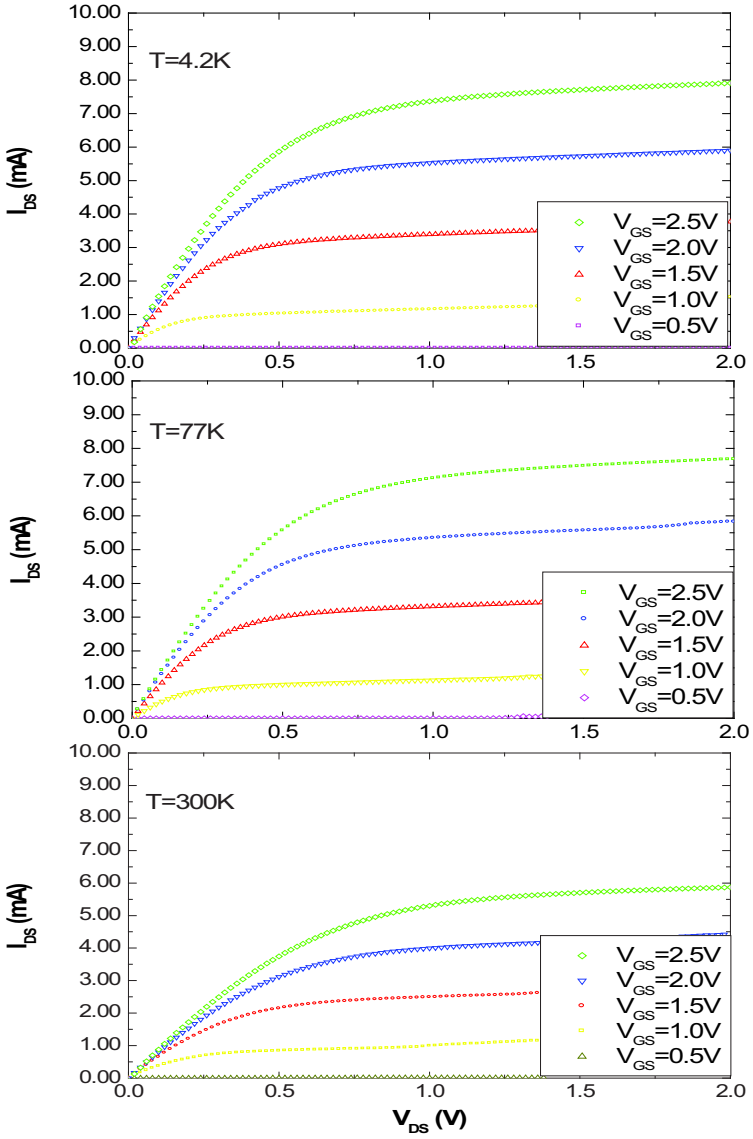


Figure 2.8: I-V curves of a short-channel NMOS device at different temperatures. The saturation current at 4 K increases by only 40% above that at room temperature due to the fact that the velocity saturation occurs early.

2.5 I-V characteristics

Both long-channel and short-channel devices were measured and the I-V curves are plotted in Fig. 2.7 and Fig. 2.8 at different temperatures. For long-channel devices, the drain current increases by a factor of 2.5 from 300 K to 4 K, benefiting from the higher low-field mobility. Note that although the low-field mobility at 4 K is more than 10 times larger than at room temperature, the high-field mobility does not benefit very much from low temperature. That explains why low-field mobility goes up more than 10 times while the drain current goes up less than 3 times. For short-channel devices, however, the applied field is higher than the saturation field and the velocity saturation limits the current performance of the devices. The drain current therefore increases only about 50% from 300 K to 4 K.

2.6 Subthreshold current

The basic assumption of the MOSFET analysis is that no inversion layer charge exists below the threshold voltage, so there is zero current below threshold. The actual subthreshold current is not zero but reduces exponentially below the threshold voltage. The subthreshold behavior is critical for dynamic circuits since it determines the static power and retention time of a dynamic memory cell.

The subthreshold current is caused by diffusion of inversion charge existing in weak inversion. Since drift is negligible, we can calculate the subthreshold current as:

$$I_d \propto \frac{dQ_{inv}}{dy} \approx \frac{Q_{inv}(y=0)}{L} \tag{2.13}$$

$$Q_{inv} = q\sqrt{\varepsilon_s/2qN_a\varphi_s}\frac{n_i^2}{N_a}e^{q\varphi_s/kT} \tag{2.14}$$

$$I_d = kTD_n\frac{W}{L}\sqrt{\varepsilon_s/2qN_a\varphi_s}\frac{n_i^2}{N_a}e^{q\varphi_s/kT} \tag{2.15}$$

From the above equations, we see that the subthreshold current depends heavily on the surface potential φ_s. In general, from Poisson's equation applied to a MOS capacitor, the surface potential has a relationship to the gate voltage,

$$\frac{\Delta V_g}{\Delta\varphi_s} = 1 + \frac{C_{dep}}{C_{ox}}, \tag{2.16}$$

where the C_{ox} and C_{dep} are gate oxide capacitance and depletion capacitance under the gate, which form a capacitive voltage divider. Therefore, if we define an inverse slope on a log-linear scale I-V curve ($swing$) as a performance metric, we find that this metric is temperature dependent as,

$$S \equiv (\frac{d(logI_d)}{dV_g})^{-1} = ln(10)\frac{kT}{q}(\frac{d\varphi_s}{dV_s})^{-1} = ln(10)\frac{kT}{q}(1+\frac{C_{dep}}{C_{ox}}) \tag{2.17}$$

The key point for a minimized swing is to minimize the depletion capacitance. However, the kT/q term limits the swing at certain temperatures. At room temperature, the best device one can make has a swing of 60 mV/dec (assuming zero depletion

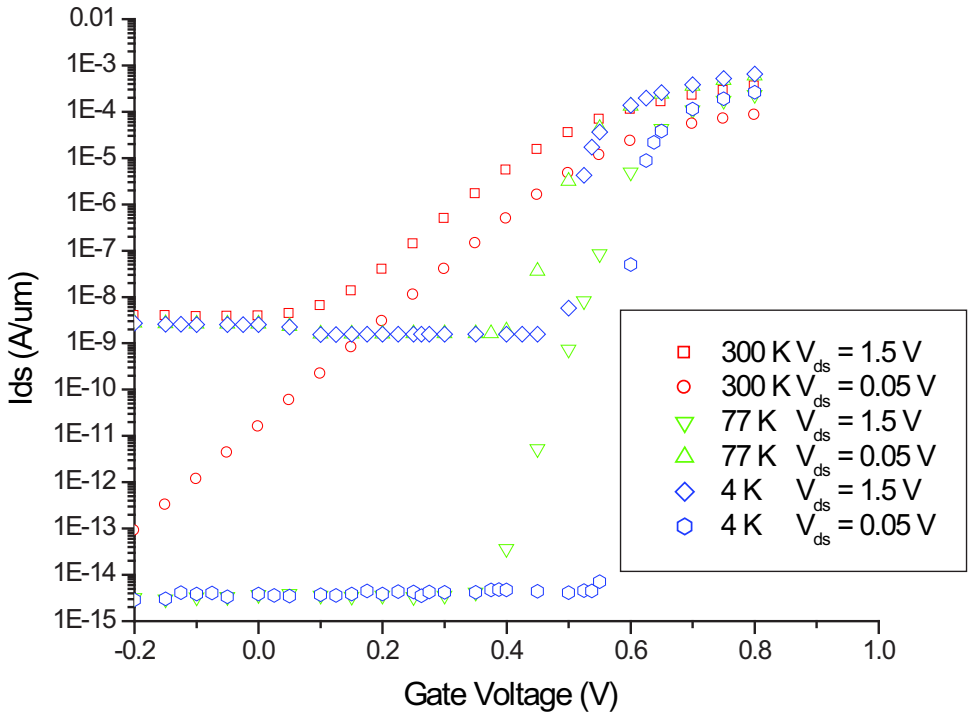


Figure 2.9: MEDICI simulation of subthreshold characteristics of an NMOS device at different temperatures.

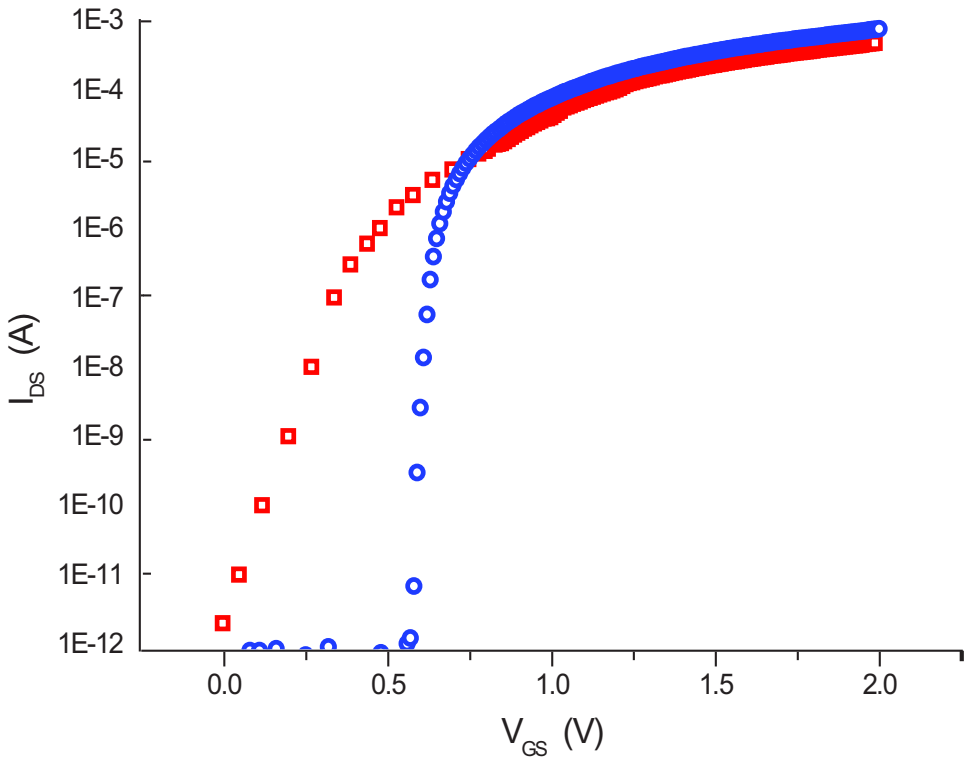


Figure 2.10: Measured subthreshold characteristic of an NMOS device at different temperatures. The subthreshold swings at 4 K and at room temperature are 9.6 mV/dec and 75 mV/dec, respectively.

capacitance), and modern processes typically end up with swing in the range of 70 mV/dec to 100 mV/dec. Devices with subthreshold swing larger than 100 mV/dec are considered as bad devices with excessive leakage current.

Subthreshold current is probably the most important limiting factor that impedes low-voltage (therefore, low-power) operation of modern CMOS circuits at room temperature, without speed performance degradation. It is especially important to keep the subthreshold current small for large-scale memory applications. For example, even 1 nA leakage current in a single memory cell gives a total 1 A leakage current for a one gigabit memory system! Many research efforts have been made in order to minimize

the subthreshold current, both device-wise and circuit-wise. [38][39] However hard people work toward a smaller subthreshold swing, the kT/q term sets a limit. However, low-temperature operation can easily decrease this number to a much smaller one that results in essentially zero leakage current. MEDICI (a synopsys®computer program to simulate the semiconductor devices) simulation results of subthreshold behaviors for different temperatures are shown in Fig. 2.9, the simulated results confirm the analysis above. Also, experimental log-linear scale I-V curves are plotted in Fig. 2.10 for room temperature and 4 K. In our experiments, the subthreshold swing at 4 K is about 10 mV/dec.

Although this swing is small enough to suppress the leakage to zero, it is much higher than the predicted value, which is around 1 mV. One of the main reasons is that the assumption we made in the analysis that the current is mainly diffusion current is not true at 4 K. The diffusion current at 4 K is so small that the small drift current becomes dominant. Also, quantum effects and 2D electron gas (2-DEG) effects arise in the inversion layer of the MOSFET at low temperatures [40], which bring in more current components that effectively increase the subthreshold swing.

One very important effect of smaller subthreshold swing is the essentially infinite retention time of a DRAM cell. Our collaborators at Yokohama National University, measured the memory retention time at both room temperature (300 K) and liquid helium temperature (4 K) [42]. The 3-T memory cell's retention time at 300 K is about several seconds while the retention time at 4 K is more than 24 hours according to

their experiments. And they also measured retention time at different temperatures, the retention time versus temperature curve is plotted in Fig. 4.8 and the extrapolated retention time at 4 K is estimated at 10^{482} years.

2.7 MOSFET Capacitance

Capacitances of MOSFETs are of great importance for their operation, especially for digital circuits and high-frequency operation of analog circuits. Low-temperature capacitances, therefore, attract a lot of research interest. The research on MOSFET capacitances at low temperatures helps not only to build a complete low-temperature model for CMOS, but also to understand the low-temperature physics in MOSFETs.

2.7.1 Gate capacitances

The gate capacitance of a MOSFET constitutes the input capacitance and, therefore, has a large effect on the speed of the digital circuits. Fig. 2.11 is the curve of the gate capacitance measurement at various temperatures. All the measurements were done by the HP E4980A LRC meter controlled by a computer. The resolution of the LRC meter is about 1 pF, which is the reason we used a very large NMOS (W = L = 100 μm) to make sure the capacitance was in the measurable range. The room-temperature gate capacitance (as in the inversion region or accumulation region), according to the NSC's parameter table, is supposed to be about 60 pF with a

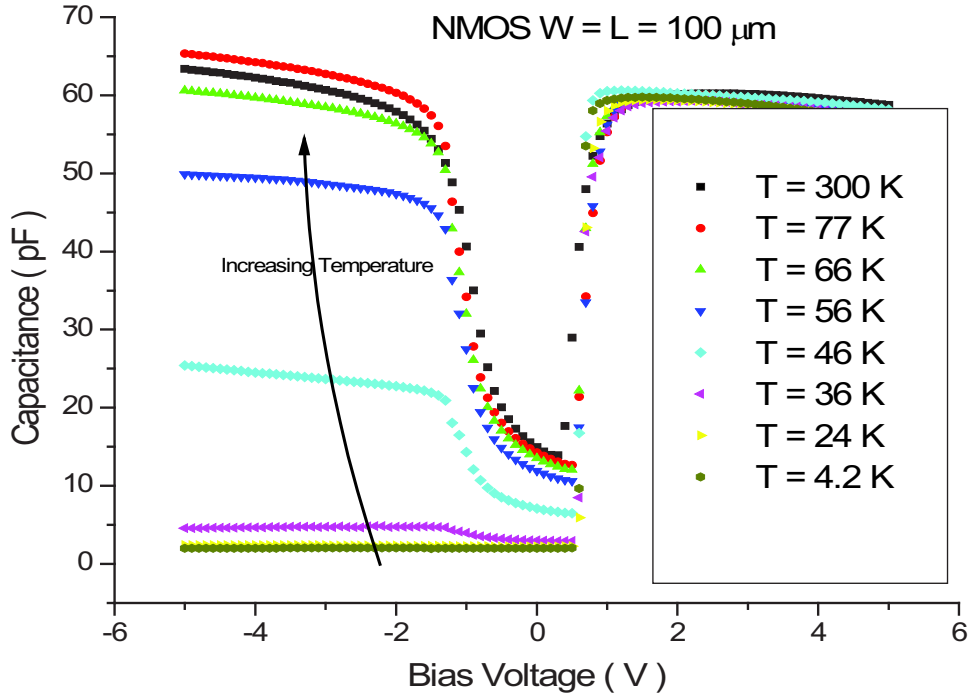


Figure 2.11: Gate capacitance of an NMOS device at various temperatures. Below 77 K the freeze-out effect dominates in the accumulation region (when $V_{gs} \leqslant V_{FB}$) and depletion region (when $V_{FB} \leqslant V_{gs} \leqslant V_T$, so the capacitances in these two regions decrease with temperature. While the capacitances in the inversion region (when $V_{gs} \geqslant V_T$) do not change with temperature.

5% variation, which fits well with our experimental results.

One can tell from the curve that in the accumulation region and depletion region, gate capacitance decreases with temperature when the temperature is below 77 K, and there is little difference of gate capacitance in the inversion region for different temperatures. The physics behind the phenomena can be explained as follows. In the accumulation region, the mechanism of gate capacitance is that the holes are generated by the thermal excitation right at the surface of the interface between silicon bulk and oxide, the capacitance then is equal to the oxide parallel-plate capacitance. However, when temperature drops, the thermal excitation gets weaker and weaker, and field-induced ionization is very weak in the bulk region. Fewer and fewer holes are generated until the number of holes cannot respond to the voltage variation, which means the capacitance gets smaller. In the depletion region, the capacitance is the parallel-plate capacitance in series with the depletion capacitance. When the temperature drops, the depletion width increases (due to the surface-potential increase as the temperature drops). Therefore the total capacitance decreases. As to the inversion region, the mechanism of the capacitance is the inversion carriers' movement. The inversion carriers, electrons, are collected from the heavily doped source and drain regions in spite of the change of the temperature. So the temperature does not affect the capacitance in inversion region.

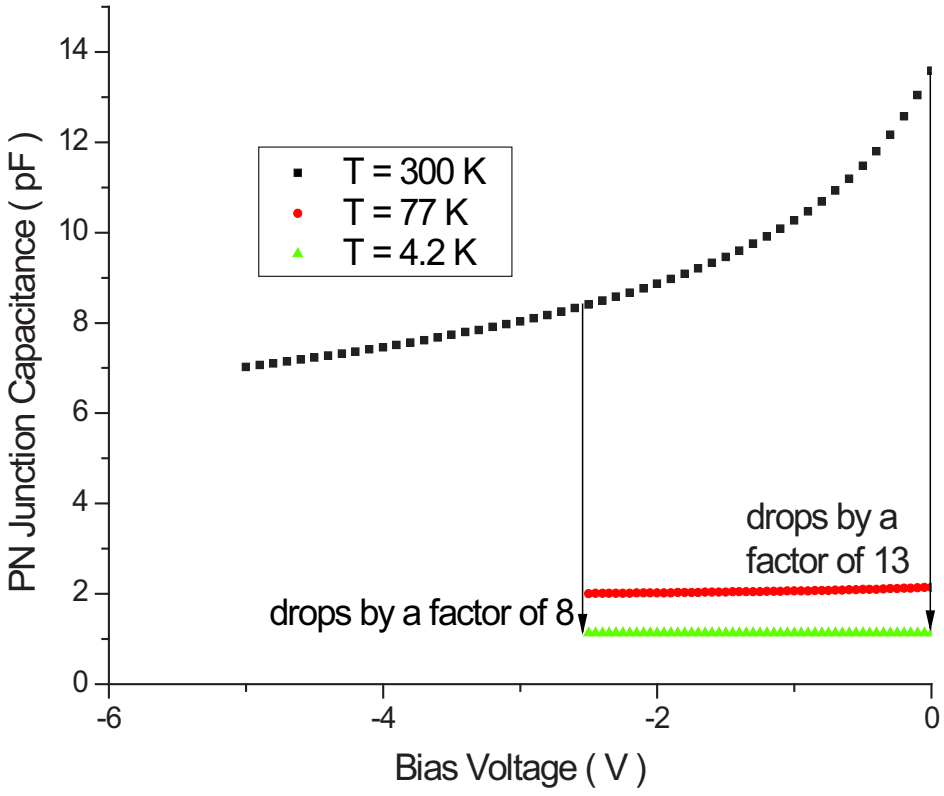


Figure 2.12: Drain/source capacitance of an NMOS at different temperatures. The 4 K values are the minimal capacitance the machine can measure. So the 4 K capacitances in this NMOS are smaller than 1 pF. It is reasonable to assume a 10 $\times$ reduction.

2.7.2 Drain and source capacitance

Drain and source capacitances are typically the circuit output parasitic capacitances. They are PN diode capacitances, which show bias-voltage dependence. The PN diode capacitance is determined by the depletion width, and the depletion width is increased when temperature drops. For NSC's 0.25 μm process, the MOSFETs' source and drain are so heavily doped that they don't freeze out even at 4 K, so their low-temperature capacitance is basically determined by the low-temperature properties of the substrate. Fig. 2.12 shows the C-V curves of a PN junction at different temperature, 4 K, 77 K, and 300 K, respectively. The experimental results fit the theoretical results ($CV^2 = \text{constant}$). Since 77 K is still in the weak frozen-out region, the substrate at 77 K is still acting as a conductor. At 4 K, however, the substrate is totally frozen out and the capacitance is therefore extremely low. All the flux lines are looking for paths to real ground, so the capacitance is no longer the depletion capacitance. Instead, the capacitance is the capacitance between the drain/source and the closest ground.

For CMOS digital circuits, gates are mostly biased at positive voltages so the gate capacitance at 4 K is not different from the room-temperature value. Drain/source capacitances, on the other hand, are at least 10 times smaller than the room-temperature values. And for a typical modern CMOS process, at room temperature, the gate capacitance is about the same as the drain/source capacitance. Therefore, the 4 K total capacitance is about half of the room-temperature total capacitance, assuming

Table 2.1: The most important model parameters at room temperature and at 4 K

Model parameters	300 K value	4 K value
VTH0 (Threshold voltage)	0.329	0.508
U0 (Mobility)	331	802
vsat (saturation velocity)	133,450	180,541
cj (junction capacitance)	2.040547e-10	2.040547e-11
RDSW (series resistance)	139	145

a fan-out of 1.

2.8 A complete BSIM model for 4 K CMOS digital circuits

In order to simulate our CMOS circuits, a 4 K CMOS model is necessary, because no commercial simulators allow setting temperature down to 4 K. (Most of them allow temperature down to 100 K.) The solution is to modify the room temperature BSIM-3 model file according to our experimental data. By doing this, the simulator will treat low-temperature CMOS circuits as room-temperature ones with a different model. We established complete BSIM-3 model on this basis. Important parameters

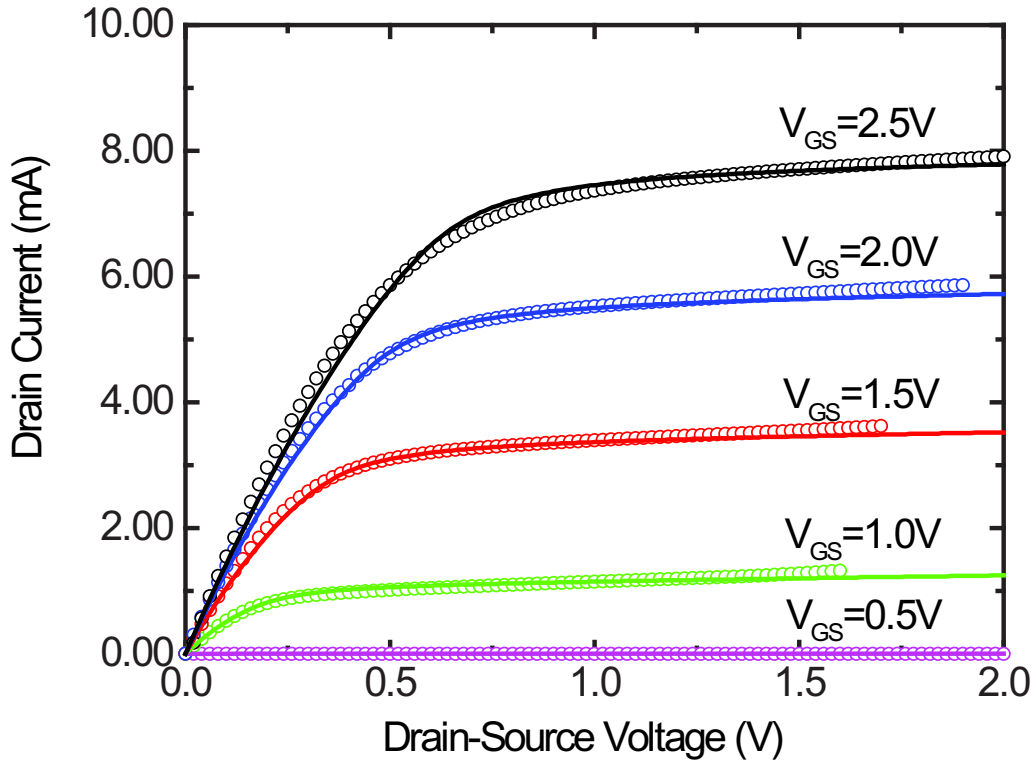


Figure 2.13: Measured and simulated I-V curves of an NMOS device at 4 K. The dots are measured results and the lines are the simulation results based on the new 4 K model. Although there are some mismatches, for digital circuit simulation, the simulated curves fit the measured ones very well.

are changed accordingly as shown in Table 2.1.

Among those parameters, the threshold voltage, the capacitance values, and the velocity saturation are based on experiments. The series resistance value is not based on experiments. Instead, it is extracted by fitting the measured I-V curves with simulated ones. The simulated and measured I-V curves are plotted in Fig. 2.13.

I-V curves are not enough to verify the new 4 K model. An inverter-based ring oscillator circuit was designed and tested at both room temperature and 4 K, and simulations based on the room-temperature model as well as the 4 K model were

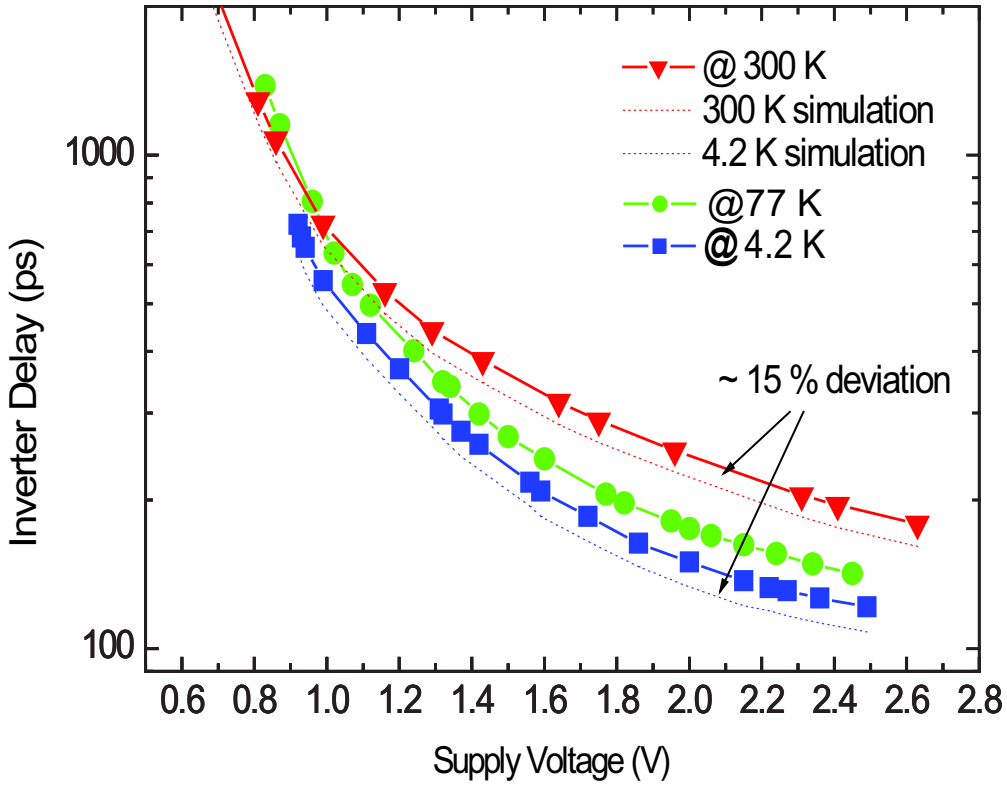


Figure 2.14: Ring-oscillator measurements and simulation results at different temperatures. The simulation results at room-temperature and at 4 K are based on the room-temperature model and the newly developed 4 K model. The 15% deviation is considered as fairly good agreement, even at 300 K.

carried out. The ring oscillator consists of 30 identical inverters and one NAND gate in series; one of the inputs of the NAND gate is connected with the output of the last inverter and the other input of the NAND gate is left as the trigger input of the oscillator. When the trigger input is high, the ring oscillator starts to oscillate with a frequency controlled by the supply voltage. The inverter versus supply voltage curve is shown in Fig. 2.14. The 4 K operation increases the speed of the ring oscillator by about 50%, according to both the simulations and the measurements. The simulation results based on the new 4 K model fit well with the measurements. Since inverters are the most elementary logic circuits, we can conclude that the new 4 K CMOS model works well for digital circuit design.

In digital circuit design, in order to achieve the best speed performance for a large and complicated circuit, the "logical effort" approach [41] has to be used. The basic idea is to minimize the total delay by sizing every stage of the circuit such that the delay for each stage is the same. For 4 K CMOS digital circuits, this approach is still valid. The only difference is that the ratio of "self-load" capacitance (drain/source capacitance) to the intrinsic capacitance (gate capacitance) at 4 K is almost zero while the room temperature value is about unity for a typical modern CMOS process. The consequence is that the optimum fan-out for 4 K CMOS circuits is 2.78 rather than 3.6 for room-temperature CMOS circuis. All design procedures remain the same.

One more comment about our 4 K CMOS model is that the model works fine for digital circuits, but not well for analog circuits. In analog circuit design, important

parameters are $r_o = (\frac{dI_{ds}}{dV_{ds}})^{-1}, g_m = \frac{dI_{ds}}{dV_{gs}}$, which are differential parameters. Although our model can fit I-V curves rather well, it is hard to make it fit those differential parameters. For analog circuits at 4 K, very few published designs are available and most of published results are based on empirical design. [43]

2.9 Power consumption of CMOS digital circuits at 4 K

Power consumption of digital circuits is becoming more and more important, as discussed in Chapter 1. As the technology is scaled to below 100 nm, the power density problem is increasly severe. Power consumption of a digital circuit consists of dynamic power and static power. The dynamic power is the power consumed on charging capacitors, and can be written as $P_{dyn} = CfV_{DD}^2$; the static power depends on the subthreshold leakage current and can be written as $P_{st} = I_{leak}V_{DD}$. In long-channel devices at room temperature, since the leakage current is small and most of the power is dynamic. However, for a typical modern short-channel CMOS process, the subthreshold leakage current cannot be ignored. The static power can be higher than 30% of the total power, which is a critical problem for modern CMOS digital circuits. The 4 K operation, however, can easily eliminate the 30% static power because of the sharp subthreshold slope.

Consider as an example, a digital system having at 300 K 30% static power and

70% dynamic power consumption cooled to 4 K with the same supply voltage, and assuming the speed is boosted by 50%. The total power consumed at 4 K can be written as

$$\begin{aligned} P_{4K} &= C_{4K} f_{4K} V^2 + P_{st,4K} = \frac{C_{300K}}{2} f_{300k} \times 1.5 \times V^2 + 0 \\ &= \frac{1.5}{2} \times 0.7 P_{300K} \approx \frac{1}{2} P_{300K} \end{aligned} \tag{2.18}$$

So the 4 K operation can lower the total power consumption by a factor of 2, even at a 50% higher speed.

2.10 Special low-temperature CMOS processes targeting low-voltage high-speed operation

We have shown that 4 K digital circuits using commercial CMOS process can achieve better speed performance and lower power consumption. More improvements can be obtained if the CMOS process is specially designed for 4 K operation.

In order to achieve optimized threshold voltage, one has to trade off between subthreshold leakage and speed. To get reasonably controlled leakage, the threshold voltage should not be less than three times the subthreshold swing; to get reasonable speed, the supply voltage should not be less than four times threshold voltage. If we follow these two rules to design the 4 K CMOS devices, the threshold voltage can be as low as 30 mV, and the supply voltage can be as low as 120 mV. Because of

the 10 times higher low-field mobility, 120 mV supply voltage does not degrade the speed performance very much. Considering only the quadratic relationship between dynamic power and supply voltage, this low-voltage 4 K CMOS process would decrease the dynamic power by a factor of more than 100, compared with a regular room-temperature CMOS process! And the speed would remain almost the same!

However, there are two main limitations on the threshold voltages. The first limitation is the channel punchthrough effect. The main approach for a smaller threshold voltage is to decrease the channel doping level, according to Eq. 2.4. However, if the channel doping level is too small, the depletion width of the junctions between the source/drain and the body will be large enough that the two depletion regions will merge together, causing a large drain-source current at higher drain-source voltage. This effect is called punchthrough and can only be prevented by increasing the channel doping level. The second limitation for small threshold voltage is the spread of threshold voltages. In modern CMOS processes, the spread of threshold voltages at room temperature can be as high as 20% and most contributing factors are not temperature scalable. For example, one important factor that causes threshold spread is trapped charges. Those charge sitting inside the silicon oxide between gate material and channel silicon. They contributes Q_{trap}/C_{trap} to the total threshold voltage. The spread of these charges contribute to the spread of threshold voltage. Since these charges do not go away or decrease in value at low temperatures, the 4 K operation does not reduce the spread caused by these charges.

Table 2.2: Comparison of room-temperature CMOS and 4 K CMOS

Metrics	300 K CMOS	4 K CMOS	Special 4 K CMOS
Speed	Fast	Faster	Faster
Robustness	Good	Better	Better
Power consumption	Low	Lower	Much lower
Reliability	Good	Good	Good

Even with these limitations, it is still possible to make the threshold voltage as low as 100 mV and supply voltage as low as 400 mV without having punchthrough in channels or having unacceptably high threshold variations. This newly designed CMOS process would bring a power consumption improvement by one order of magnitude, compared with the regular room-temperature CMOS process.

2.11 Conclusion

Low-temperature operation of commercial room-temperature CMOS not only functions well, but also boosts the performance of digital circuits by providing more on-current, essentially zero off-current, and smaller parasitic capacitances. We have studied the devices and circuit characteristics of a commercial CMOS process, with a

complete 4 K model. A specially designed 4 K CMOS process with smaller threshold voltage and supply voltage can further improve the performance, especially the power consumption. Table 2.2 compares the three situations. Of course, the specially designed 4 K CMOS is the best among those three, but the commercial CMOS operated at 4 K is good enough for proof of the concept of the hybrid memory system. All designs and simulations in the next chapter are based on commercial CMOS operated at 4 K. And we will lay down a foundation for the design of a hybrid memory system.

Chapter 3

Design and simulation of a hybrid memory system

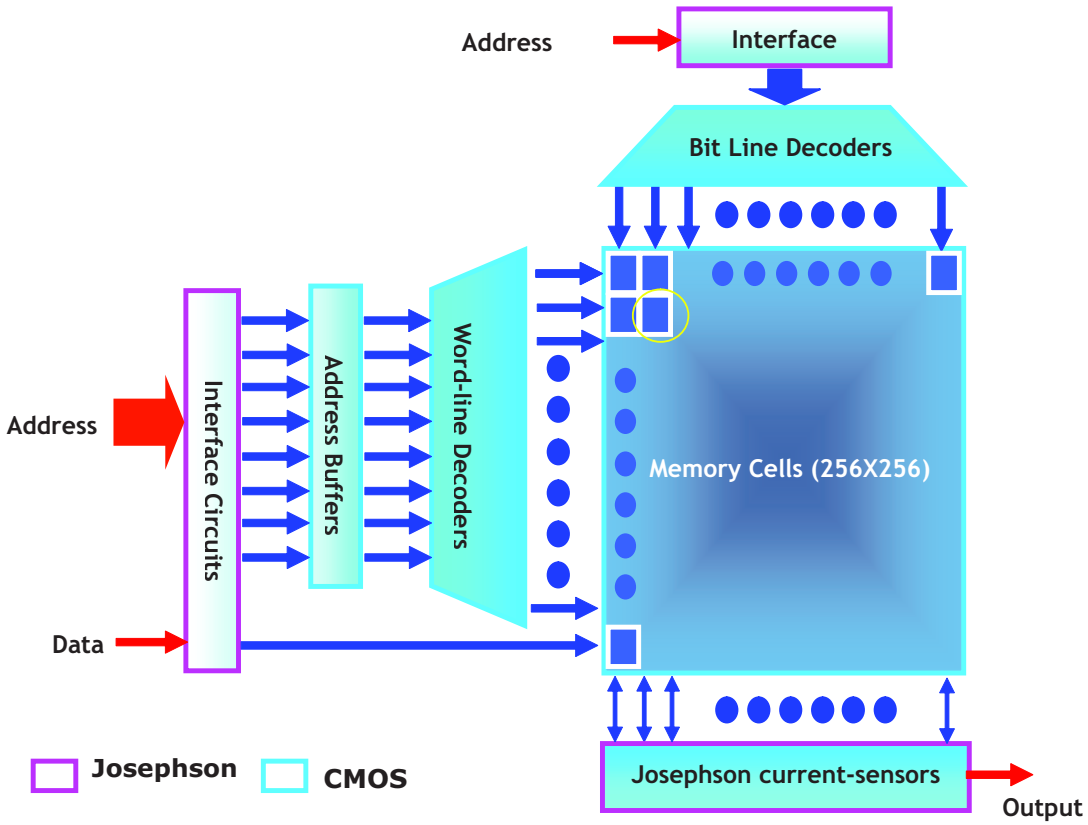


Figure 3.1: The system block diagram of a 64-kb Josephson-CMOS hybrid memory system. The memory core and decoder are fabricated in a commercial CMOS technology, the current sensors are fabricated by a standard Nb technology, and the interface circuits involve both technologies. The memory cell is the traditional 3-transistor DRAM cell, which works as a static memory cell at 4 K due to zero subthreshold leakage currents.

3.1 Overview of Hybrid Memories

Due to the density and yield problems, there is so far no successful superconducting memory larger than 4 kb [44]. The Josephson-CMOS hybrid memory idea was proposed [17] to solve the problem. The main idea of "hybrid" is to use high-density charge-storage MOS memory cells and access them by high-speed low-power superconductive devices, which takes advantage of the best features of each technology.

Fig. 3.1 shows the system block diagram of a 64 kb hybrid memory. As shown in Fig. 3.1, in a Josephson-CMOS hybrid memory, data storage and address decoding are implemented in CMOS technology, and the CMOS memory communicates with the superconducting CPU via input interface circuits and current sensor outputs. An interface circuit transforms SFQ pulses that come from a superconducting CPU to CMOS digital signals. A current sensor carries the same responsibility as a sensor amplifier in a semiconductor memory system, sensing the data stored in the memory with the output being an SFQ pulse.

In order to solve the long-standing memory bottleneck faced by superconducting digital electronics, the proposed hybrid memory must have advantages over Josephson memories in terms of capacity density, robustness, and speed. For a hybrid memory, the CMOS part is extremely dense, as we discussed before, and is scalable. Notice that the area of memory cells is the dominant part, hybrid memory wins in terms of density. The CMOS logic family is well known for its robustness; the interface circuit and the current sensor are not as strong as CMOS digital circuits, however, with careful design, we can achieve acceptable robustness. And the robustness of the hybrid memory is higher than Josephson memories. As for speed, although CMOS logic is slower than SFQ logic, CMOS technology is still advancing and following Moore's Law but the SFQ logic has already approached the quantum limit due to the intrinsic properties of the SFQ logic. Besides, for a semiconductor-based computing system, it is normal that the speed of memories is much lower than the speed of

processors anyway. However, as will be shown in the next few paragraphs, the speed of the hybrid memory can be faster than that of a room temperature CMOS memory, due to not only the low temperature operation, but also the current sensing scheme. And the speed of the hybrid memory can be as fast as Josephson memories.

Based on the comparison, the Josphson-CMOS hybrid memory is believed to be the most promising solution for the long-standing problem. This chapter will focus on the concept, design, simulations, and optimization of a 64-kb hybrid memory system, and lay down a foundation for larger sizes of hybrid memories using more advanced technologies and working at higher frequencies.

3.2 Suszuki stack

The most challenging part in the hybrid memory system is the interface from SFQ pulses to CMOS logic signals. Typically, an SFQ pulse is a voltage pulse whose integral over time is a magnetic flux quantum ($\Phi_0 = h/2e = 2.07 \times 10^{-15} Wb$), with a typical height of 1 mV and duration of 2 ps. And CMOS logic signals are voltages in the order of 1 V, typically with a subnanosecond cycle time. The primary goal of an interface circuit is to transform an SFQ pulse into a CMOS digital signal in a timely and low-power manner. In order to keep the access time substantially less than 1 ns for a 64-kb memory system, the delay of this interface circuit has to be on the order of 100 ps or less. Because the power dissipation of the hybrid interface circuits turns out to be the dominant part of the total system power, it is critical to minimize the

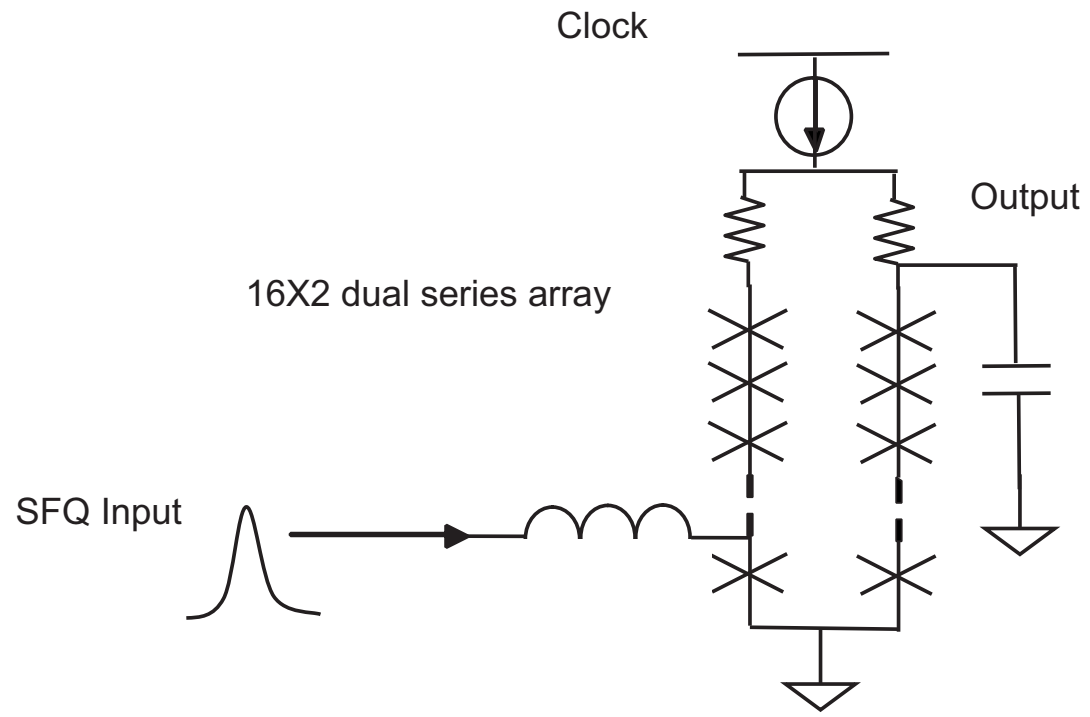


Figure 3.2: A Suzuki stack with an inductor at its front-end. The inductor transforms an SFQ pulse into a current step feeding the Suzuki stack. The bias current is synchronized with the CMOS clock. The current source can be implemented by resistors or a MOSFET working at subthreshold region.

power without adding too much delay time. Last but not least, the robustness of an interface circuit must be sufficient to realize large scale hybrid memories.

The interface circuit consists of two parts, the preamplifier and second-stage amplifier. The preamplifier works as a translator from SFQ pulse signals to 40 mV clocked signals, while the second-stage amplifier completes the amplification. A widely used preamplifier was proposed by H. Suzuki and is often called a Suzuki stack [45]; it has been extensively studied and has proved to be a strong candidate for SFQ-to-voltage-state logic conversion. [46]

Fig. 3.2 shows the schematic of a Suzuki stack with a inductor and a junction as its front-end. The operation of the circuit follows. The two parallel stacks are biased by a clock signal which is an attenuated CMOS clock signal. When the clock is high, the junction arrays are biased close to the critical current of the junctions; a typical value of the junction bias current is $0.8I_c$. The bias margin depends on frequency and process spread and will be addressed later. When the input signal arrives, the inductor transforms the input SFQ pulse into a current (Φ_0/L) feeding into the left bottom junction of the Suzuki stack and therefore switching that junction. That junction goes into a high-resistance state causing most of the current that flowed in the left branch to transfer into the right one, causing all junctions in the right branch to switch. The switching of the right-branch junctions destroys the current balance again and pours some current back to the left branch causing all other junctions to switch. After the current balance is restored, the current in the junctions will be $0.8I_c$ again but all junctions are now in the voltage state rather than superconducting state, delivering an NV_g voltage at the output node, where V_g is the gap voltage of the superconductor, typically 2.6 mV for a good quality Nb junction. After this switching, all the junctions will experience Josephson oscillations to support such a gap voltage and will not automatically reset to superconducting state as long as the bias clock signal is high. After the bias clock is turned off, all the junctions will be reset to the superconducting state, and the output goes back to zero. Since the clock signal is synchronized with the CMOS clock, the SFQ pulses will be transformed to

synchronized CMOS logic signals.

3.2.1 Delay of a Suzuki stack

The delay time can be divided into two phases. During the first phase, the left bottom junction switches to the voltage state, causing part of the left branch current to flow into the right branch and this can be represented by a voltage source pumping a clockwise current in the 2N-junction-and-2-resistor loop. So this part of delay consists of the delay of one junction switching and the current redistribution. The second phase of the delay is the delay of the switching of 16 junctions in the right branch.

The current redistribution between the left and right branches, on the other hand, is even faster than the junction switching, based on simulations using a 2.5 kA/cm^2 Nb process. The current redistribution process can be simply modeled as current redistribution in a RLC loop, where L represents all the parasitic inductance including the nonlinear junction inductances; C represents the output capacitance to ground, including parasitic capacitance and junction capacitance in series; and R represents the resistance of the two physical resistors in the loop. The analytical solution of this second-order system is an oscillation with a damping factor. The time constant associated with the oscillation is $\sqrt{LC}$ and the time constant associated with damping is L/R. What we care about here is the delay from when the step event occurs to the time the current pumping into the inductor goes beyond the critical current of the

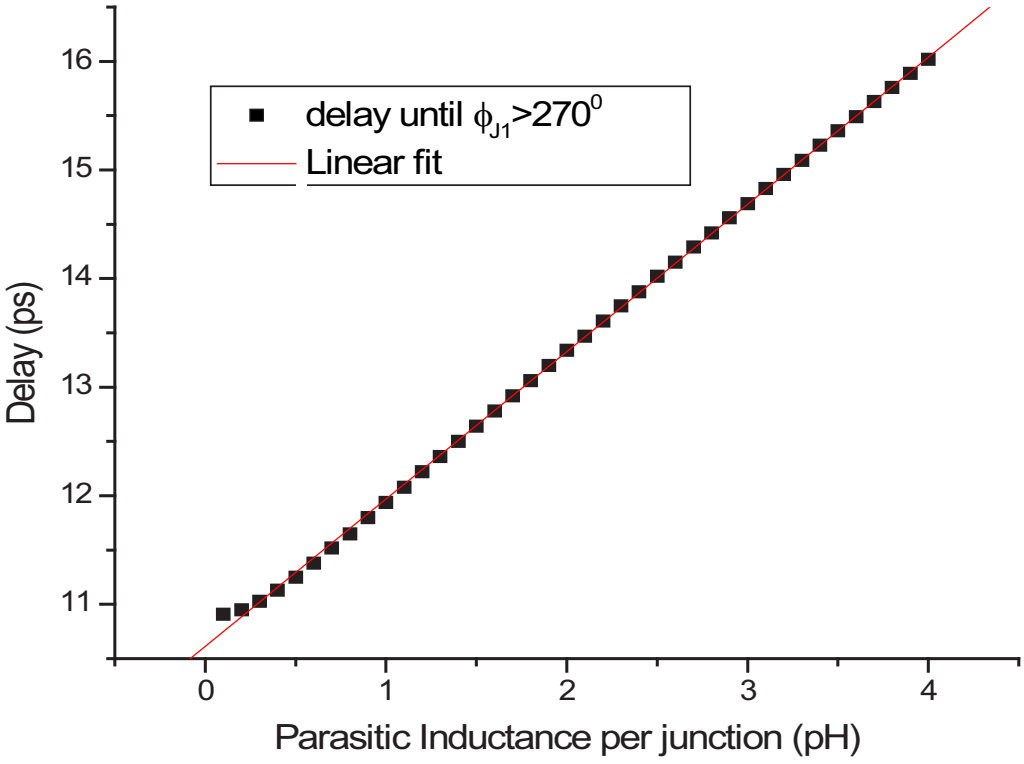


Figure 3.3: The relationship between switching time and the parasitic inductance in a Suzuki stack simulation. This result shows that the delay time is linearly proportional to the inductance, which verifies the first-order analysis.

junctions. According to the analysis of this second-order system, the delay is L/R for small capacitance. For a 2×16 junction dual array Suzuki stack fabricated by a 2.5 kA/cm^2 Nb process, the inductance of each junction is about 0.7 pH and the capacitance of each junction is about 0.15 fF, after the ground plane underneath the Suzuki stack is removed (in order to minimize the capacitance; more details are discussed in Sec. 3.4). Fig. 3.3 shows the relation between the delay for current transfer to the right leg and the parasitic inductance. The delay from simulations is defined as the delay between when the input current arrives and the phase of right branch junctions arrives at $3\pi/2$ (including the switching of left bottom junction). The straight fitting line shows that the delay of the current redistribution is proportional to the inductance, which confirms the analysis above.

The next phase is also a charging process. After the current in the right branch reaches the critical current of the junctions, all the right branch junctions will switch, giving an NV_g voltage at the output node. Here we must take into account the load capacitance contribution to the delay. The charging time is

$$t_{charge} = \frac{NV_g(C_L + C_J/N)}{I_c} = t_J + \frac{NV_gC_L}{I_c} \tag{3.1}$$

Therefore, the total delay for the Suzuki stack is

$$t_{Suzuki} = \frac{V_gC_J}{I_c} + \frac{L}{2R + R_J} + \frac{NV_g(C_L + C_J/N)}{I_c} = 2t_J + \frac{L}{2R + R_J} + \frac{NV_gC_L}{I_c} \tag{3.2}$$

For this dissertation, the superconducting chips were fabricated by the Superconductive Research Laboratory in a facility formerly a part of NEC (NEC-SRL) under

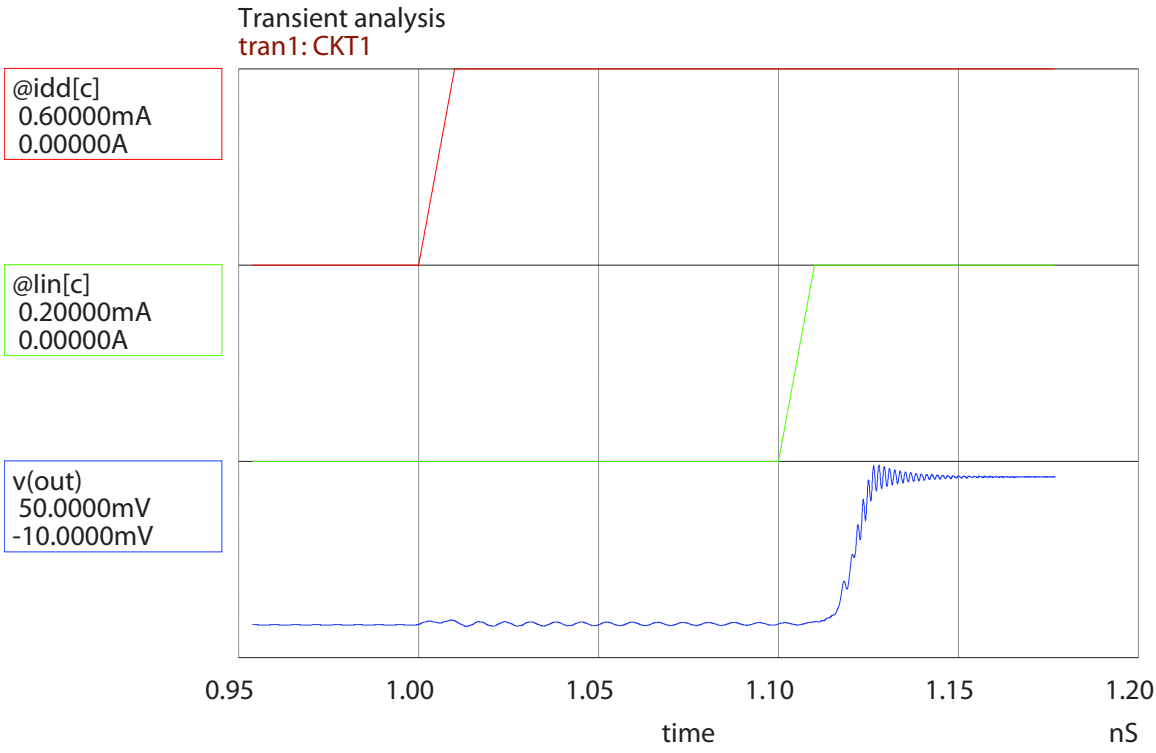


Figure 3.4: The WRSPICE simulation results of a 2×16 JJ Suzuki stack using a 2.5 kA/cm^2 Nb process with a 10 fF capacitive load. The total turn-on delay is about 20 ps.

a joint program with National Yokohama University, Japan, using a standard 2.5 kA/cm^2 Nb/AlO$_x$/Nb process. The Suzuki stack consists of a 2×16 junction dual stack with junction critical current of 400 μA critical current, the total delay should be less than 20 ps without load capacitance. The load capacitance, however, consists of parasitic capacitance to ground of the junction array and the input capacitance of the next stage. The parasitic capacitance of the junction array can be minimized by removing the ground plane underneath, see Sec. 3.4 for further discussion. The output of the Suzuki stack is connected to the gate of a MOSFET through a large capacitor (10 pF). The gate capacitance of the MOSFET depends on the size of the device. The total capacitance loaded oo the Suzuki stack, can be kept well below 100 fF. Fig. 3.4 shows the simulation of a Suzuki stack loaded with a 10 fF capacitor using 2.5 kA/cm^2 Nb process; it shows that the turn-on delay of the Suzuki stack contributes very little to the system delay. Even with a 100 fF load capacitance, which is a large number for a MOSFET gate and the parasitic capacitance, the delay contributed by the load capacitance is only 10 ps more.

3.2.2 Resetting time of a Suzuki stack

As is well known, all latching circuits including Suzuki stacks cannot reset by themselves. That's the reason the bias is a synchronized clock signal which resets all the junctions to zero-voltage states upon falling to zero bias. Therefore, another important consideration in the design of Suzuki stacks (as well as for other latching

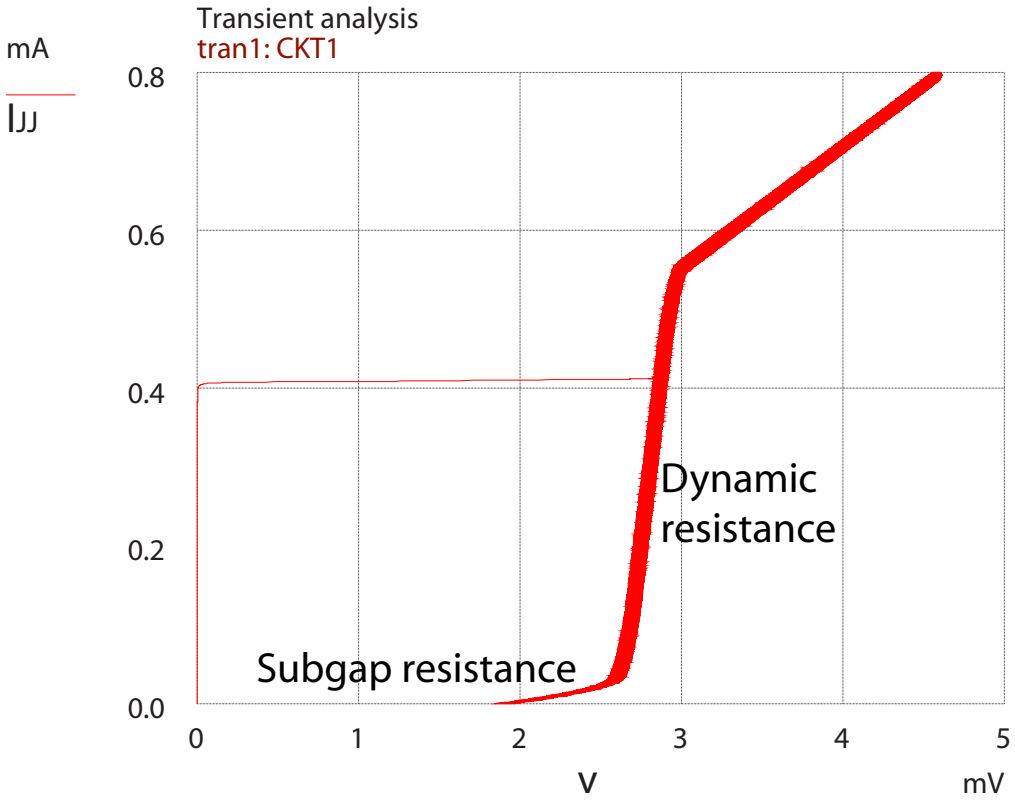


Figure 3.5: The simulation I-V curve of a single Nb junction. The dynamic resistance and the sub-gap resistance are shown in the curve. Typical values for a 2.5 kA/cm^2 Nb process are 1 Ω and 300 Ω, respectively. The wider line is due to the oscillations.

gates) is the resetting time. The resetting process is an RC discharging process with plasma oscillations caused by the nonlinear junction inductance.

The resetting time consists of two parts as follows: the first one is the discharging when the junction current is still larger than $I_{sub} = V_g/R_{sub}$. The time constant associated with this process is R_dC_J, where R_d is the dynamic resistance, shown in Fig.3.5; and the other part is the subgap region discharging time, with a time constant of $R_{sub}C_J$. The subgap resistance R_{sub} represents the quasiparticle tunneling current when the junction is in the voltage state. In an ideal junction at 0 K, this resistance

should be infinite. However, the operating temperature is typically 4.2 K and there are always pinholes in the barrier material which help to increase the quasiparticle tunneling. In a modern Nb process, the subgap resistance is on the order of hundreds of ohms for a junction with a critical current of 100 μA. The dynamic resistance, however, is typically on the order of one ohm. The subgap resistance is much larger than the dynamic resistance in a typical Josephson junction; therefore it dominates the resetting time. The resetting time therefore can be written as

$$t_{resetting} = NR_d(\frac{C_J}{N} + C_L) + NR_{sub}(\frac{C_J}{N} + C_L) \approx R_{sub}C_J + NR_{sub}C_L \tag{3.3}$$

Fig. 3.6 shows the simulation results for a resetting process, and the resetting time is around 70 ps, which fits the analysis very well. While in Fig. 3.4, the resetting time is longer in this case than that of a single junction because of the contribution of the load capacitance. And the load capacitance contributes more than it does to the switching delay.

3.2.3 Power

The power dissipation in a Suzuki stack is mainly from the quasi-static power $P = V_{bias}I_{bias}$; the dynamic power CV^2f part is about three orders less than the quasi-static power and can be ignored. The bias voltage depends on how the current source is implemented but should be at least several times NV_g. Resistor-based current sources will increase the power by a factor of 3 to 5 due to the power consumed by the bias resistor. The bias resistor cannot be too small or the after-switching current

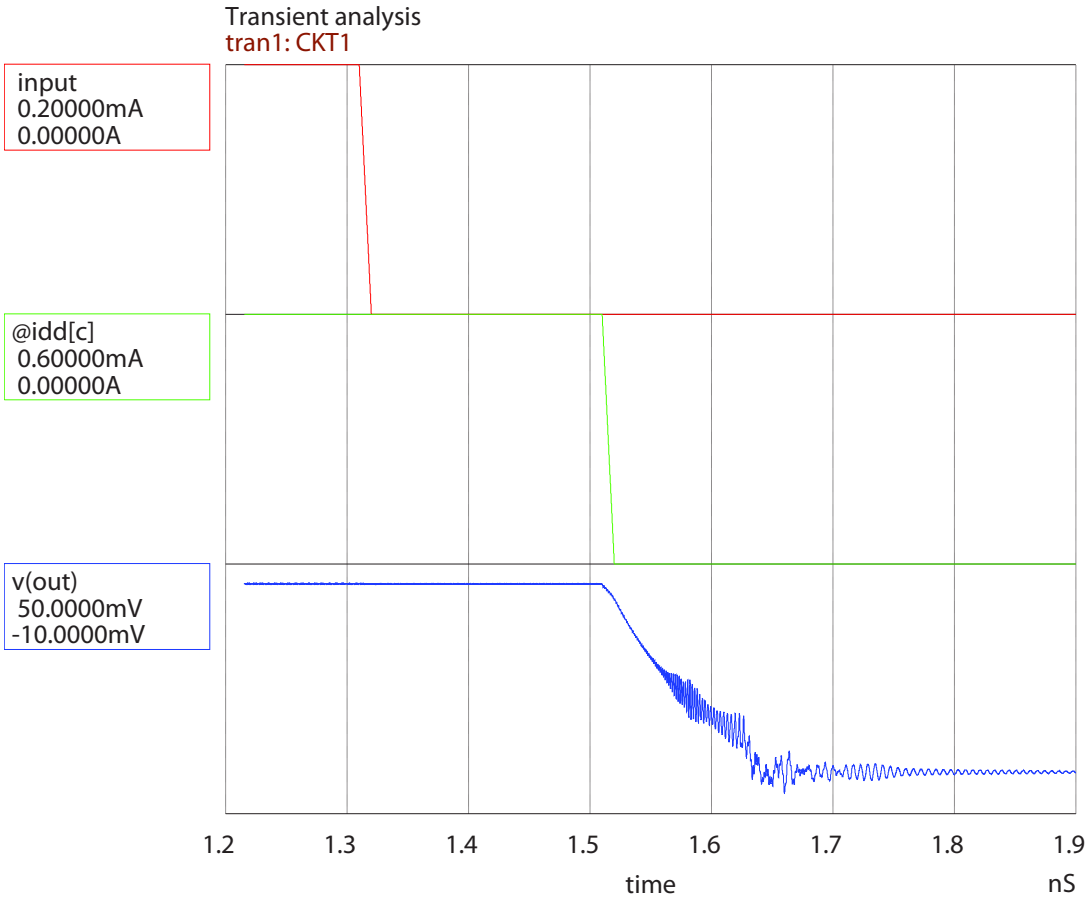


Figure 3.6: The WRSPICE simulation results for the resetting process of a Suzuki stack, with a longer delay time and complicated Josephson oscillation involved.

level cannot be sustained, with an acceptable margin and a sufficiently low bit-error rate. In practice, the bais resistor has to be larger than 3 to 5 times of NV_g/I_{bias}.

A MOSFET current source working in the subthreshold region, is a smarter choice in order to minimize the power, since the voltage drop between the drain and the source is small compared with a resistor's voltage drop. However, it also brings more complexity and reduces the yield of the hybrid due to the interconnections.

In the current design, the current source is implemented using a physical bias resistor of 200 Ω and the designed bias current is 600 μA. Therefore, the power dissipation of the Suzuki stack is 72 μW.

3.2.4 Bit-Error Rate (BER) and margins

In a high-end computation system or any communication system, bit-error rate (BER) is important to the functionality and performance of the system. A Suzuki stack is the very first stage of any system that requires conversion of RSFQ signals to CMOS signals, so the BER of Suzuki stacks is of great importance. Typically, bit errors come from circuit dynamics and noise.

Josephson-junction-based circuits have two different noise sources, the Josephson junction noise and resistor noise. Resistor noise is the so-called Johnson noise and the rms value of the noise current can be written as

$$i_{rms} = \sqrt{\frac{4kTf}{R}} \tag{3.4}$$

where k is the Boltzmann constant and f is the bandwidth. Note that the Johnson

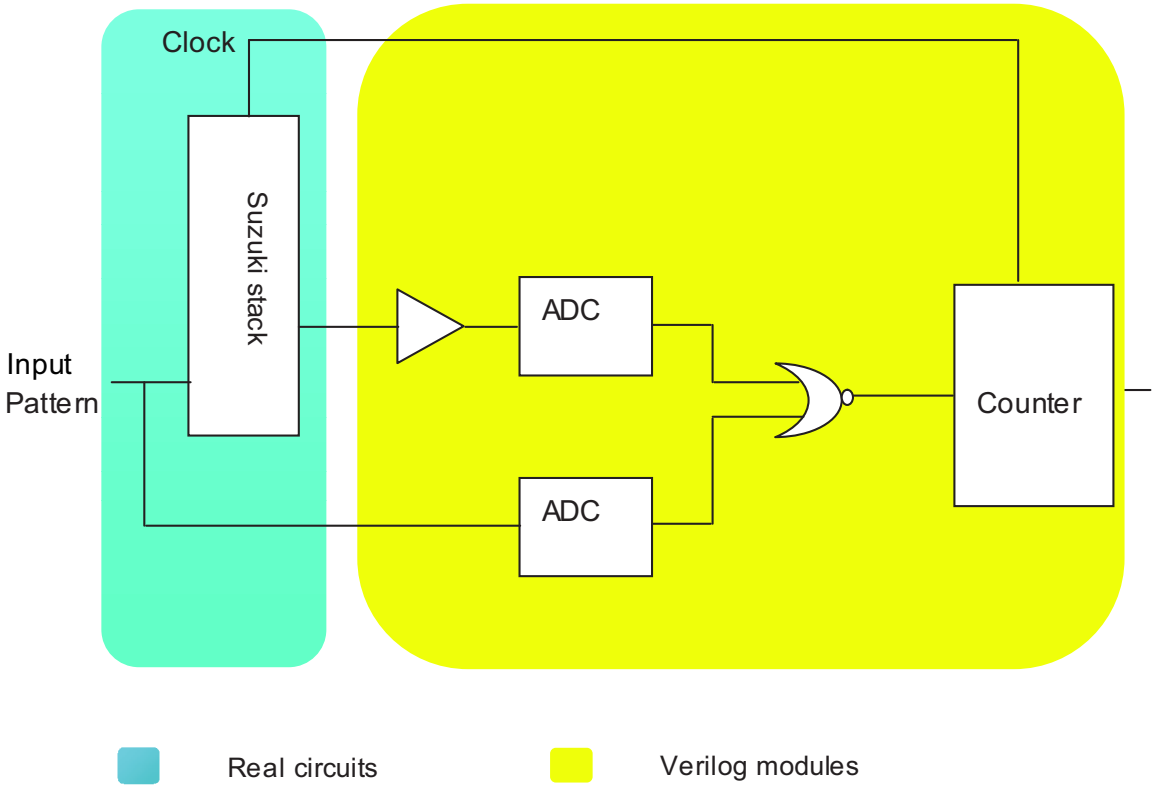


Figure 3.7: The bit-error rate (BER) simulation set-up for a Suzuki stack. The simulation is programmed to run 10,000,000 cycles, limited by the time- and memory-consuming property of this simulation. Parameters such as shunt resistance and operation frequencies are automatically changed after each simulation.

noise includes the noise generated by 4 K on-chip resistors and the noise generated by room-temperature resistors and brought down to the 4 K chip by cables.

Noise in Josephson tunnel junctions has been studied by Rogovin and Scalapino [50]. According to their study, the noise current for a tunnel junction has the form of shot noise when the voltage across the junction is higher than 1 mV, and Johnson noise with an effective resistance V/I if the voltage is smaller than 1 mV.

In order to calculate the BER of a Suzuki stack under different conditions, a mixed-signal simulation is set up as shown in Fig. 3.7. In the simulation, noises

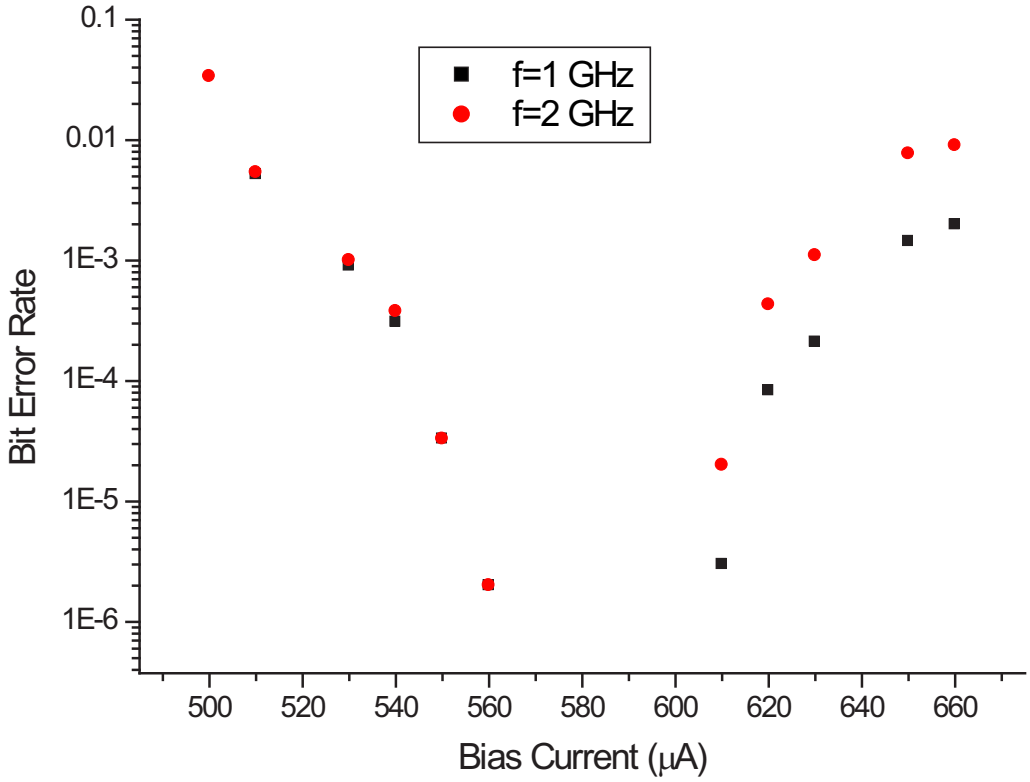


Figure 3.8: The BER simulation result of an un-optimized Susuzki stack. The junctions are unshunted and the frequencies are 1 GHz and 2 GHz. The higher working frequency leads to a larger BER when the bias current is high. When the bias current is low, however, the error rates are almost the same because there are no punchthrough errors.

are implemented by specific current sources and a $3\sigma = 5\%$ critical-current spread is taken into account. Each individual simulation runs with different initial conditions and the error events are read out by the program automatically. To keep the simulations within a reasonable time, each run was limited to 10,000,000 clock cycles; the results are plotted in Fig. 3.8. This curve makes physical sense in that there is some optimized bias current that has the best BER, and any bias current that is either larger or smaller than the optimized one will lead to more errors, depending on how far the bias current is away from the optimized value. It is easy to understand that errors caused by insufficient bias current are omission errors and larger bias currents lead to insertion errors. Because the Johnson noise dominates, the curve looks like an error function, as is predicted by the classical bit-error rate theory.

Noise is not the only parameter responsible for bit errors. The Suzuki stack is a latching circuit, it must be reset by a clock. The punchthrough effect that all latching logic circuits suffer is another source of errors, especially in high-frequency operations. The resetting time is much longer than the switching-on delay time because of the large sub-gap resistance, and there is a plasma oscillation associated with the resetting process, as shown in Fig. 3.6. In high-frequency operations, it is possible that the clock is so fast that the next input clock arrives before the previous resetting is complete. Even if the resetting process is complete and the time average junction voltage is zero, the plasma oscillation makes the situation more complicated, it is possible that the clock arrives at such a time that the oscillation helps the clock

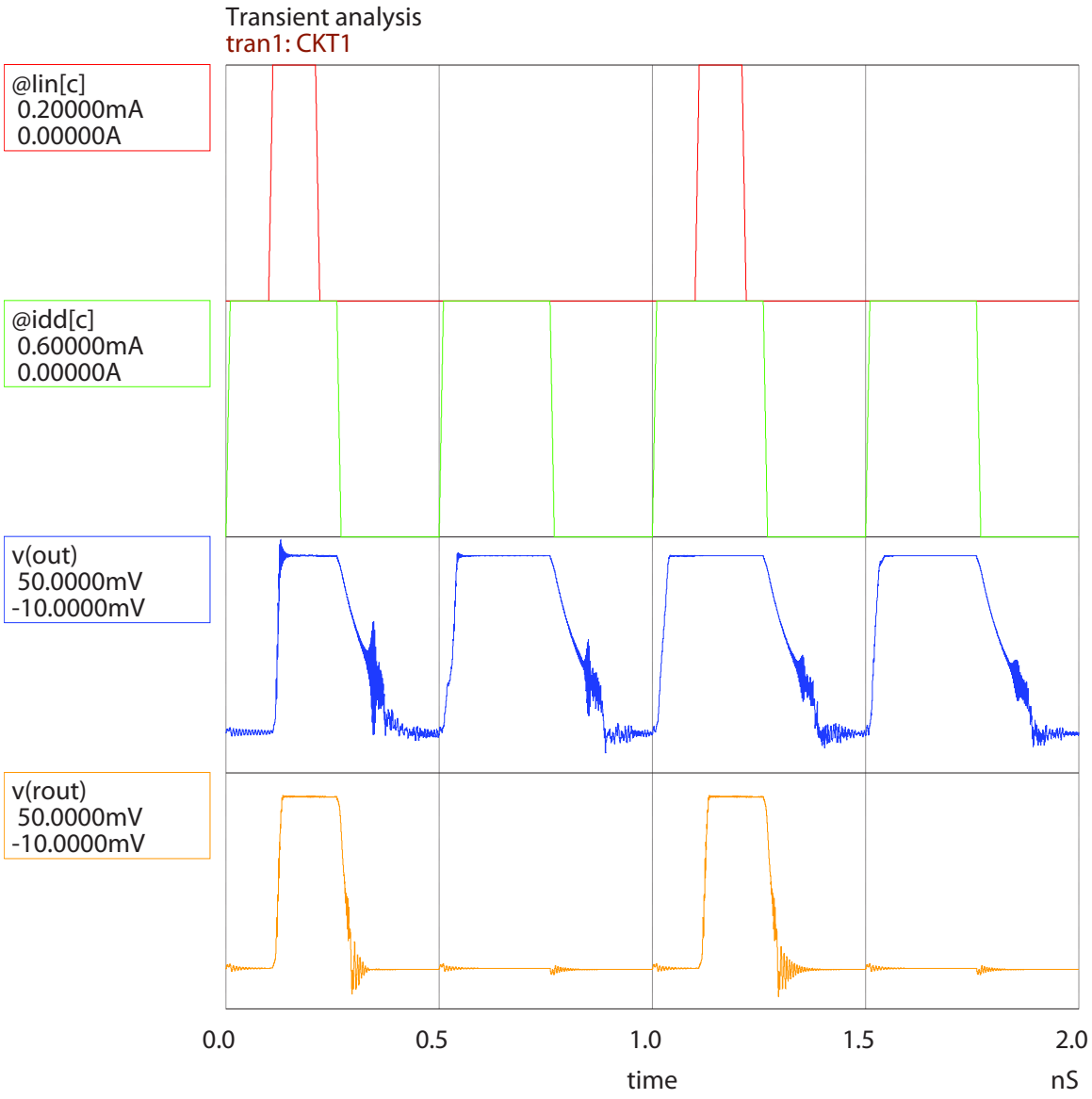


Figure 3.9: The simulation result of Susuzki stacks with and without shunting resistors on the junctions. The frequency is 2 GHz and the load capacitance is 10 fF, which makes the t_{reset}/t_{clk} ratio larger than in Fig. 3.4. The vout curve is the result of the Suzuki stack without shunted resistors and the vrout curve is the result of the Suzuki stack with 20 Ω resistors shunted on each junction.

to switch the junctions. In either way, the circuit may switch with a zero input. This is called "punchthrough" and is a source of error in any latching gate, including Suzuki stacks. In Fig. 3.9, the simulation results show two different output for two Suzuki stacks which work at a 2 GHz clock with the same 10 fF loading. One is a 2 $\times$ 16 Suzuki stack without shunted resistors and the other one is the same Suzuki stack wich 20 Ω resistors shunted on each junction. For the one without shunting, since the resetting time is larger due to the load capacitance, there are some errors caused by the punchthough effect. Due to the nonlinearity and complexity of the junction switching dynamics, the resetting waveforms are different in each clock cycle. For the one with resistors shunted, we can see from the waveform that the resetting time is dramatically reduced and, therefore, there is no punchthrough. The negative effect caused by the shunted resistors is the voltage drop, but 20 Ω resistance does not cause too much voltage drop, as shown in the waveform, which suggests it be a good way to suppress the punchthrough.

Qualitatively, the bit-error rate caused by punchthrough is a strong function of the ratio of resetting time and clock period, t_{reset}/t_{clk}; the larger the t_{reset}/t_{clk}, the higher the punchthrough possibility. However, a quantitative model is not available yet. Simulations in Fig. 3.10 also show that the BER worsens when the Suzuki stack is operated at higher frequencies, which qualitatively confirms the relationship between the BER and the t_{reset}/t_{clk} ratio.

In order to improve the BER, the first thing to do is to lower the ratio of noise

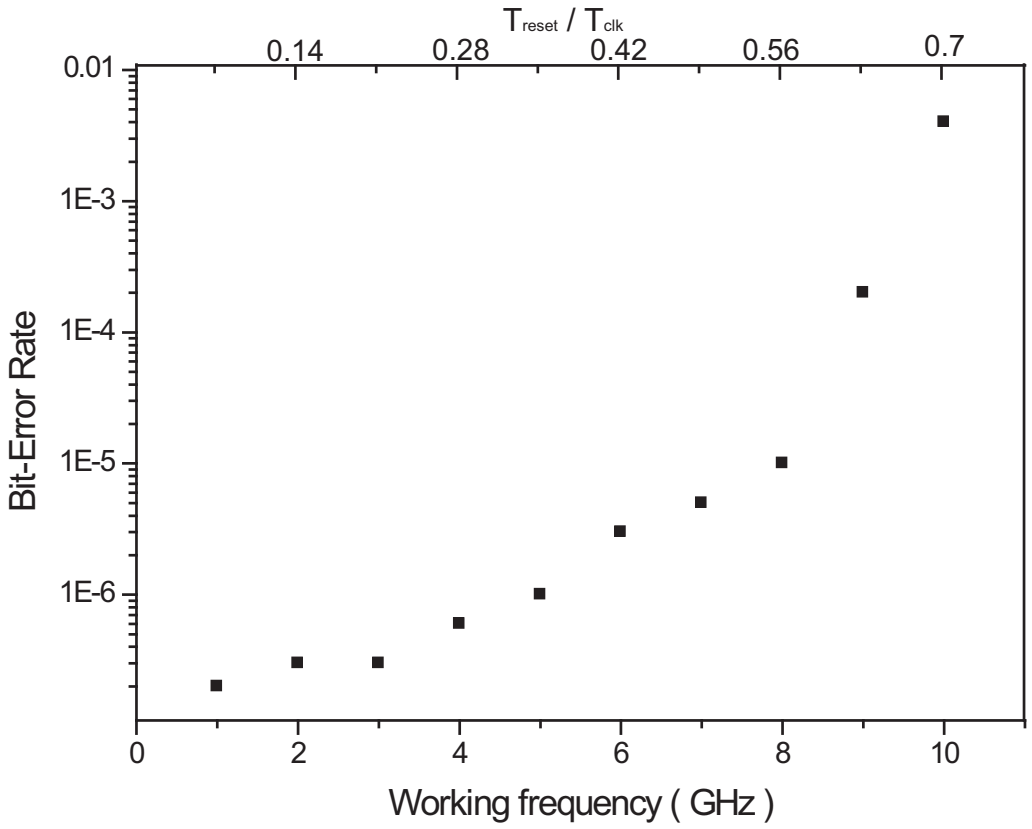


Figure 3.10: The relationship between BER and working frequencies. The current bias is 580 μA and the junctions are un-shunted. This curve confirms that the punchthrough error rate depends on the t_{reset}/t_{clk} ratio.

current to the bias current. This is accomplished by increasing the critical current of the junctions and the resistances of the two branching resistors. The second thing is to try to suppress the punchthrough effect. Adding a physical shunt resistor to each junction is a good approach to decrease the resetting time by lowering the total resistance. However, there are two issues. The first one is that is if the shunt resistors are too small, the output voltage will be reduced. The other one is that the Johnson noise introduced by the physical resistors will increase the bit-error rate. A balance between suppressing punchthrough and increasing Johnson noises has to be made in order to get an optimized value for shunt resistors for very high-frequency operation. Simulations also show that the bit-error rate depends on how fast the currents are redistributed during switching. The larger inductance helps to slow down the current redistribution and, therefore, decreases the bit-error rate.

Theoretically, a Suzuki stack has fairly large operating margins. The bias current should be at least larger than the critical current of one branch in order to make sure the stack switches when the input arrives, and less than twice the critical current in order to prevent any unwanted switching before the input arrives. This gives about a $\pm 33\%$ margin. However, many imperfections come to play in real situations. First of all, because of the process variations, not all junctions have exactly the critical current they were designed to have. A $3\sigma = 5\%$ critical current spread is typical in existing Nb processes. Secondly, physical resistors and junctions in a Suzuki stack generate thermal noise. The environment temperature is only 4.2 K, but the bandwidth is

as high as one terahertz. In this situation, a 1 Ω resistor generates a noise current with rms value 14.9 μA. Also, the room-temperature noise is brought down to low temperature superconducting chips by connecting cables. Last but not the least, the nonlinearity of Josephson junction switching dynamics will effectively make more "noise" and therefore, reduce the margins. Also, the operating margins depend on the operation frequency and the acceptable bit-error rate (BER). Quite apart from the noise effects, for high-frequency operations, the margins are lower than for low-speed situation because of circuit dynamics. The relationship between BER and margins will be addressed in the following section.

3.2.5 Optimization

The design of Suzuki stacks involves such parameters as critical current, parasitic inductance and capacitance, branching resistance, and shunt resistance. The criteria for a good Suzuki stack include turn-on delay, resetting time, power consumption, dc margin, and bit-error rate. It is very difficult to optimize the design mathematically to fit all criteria since weighting factors are hard to assign. Decisions have to be made and there are many trade-offs to consider. First of all, since they are small, the turn-on delay time and the power dissipation are not first priority. In the hybrid memory project, the Suzuki stack is only one part of the system, its delay and power are not the dominant parts of the whole system performance, as will be addressed later. The 20 ps dealy is small compared with the whole sub-nanosecond access

time, and the 72 μW power dissipation is negligible compared with the milliwatt-level second-stage power. Therefore, if we have to sacrifice delay and/or power to get a better BER, it is a worthy choice. Secondly, resetting time is an important parameter which also determines the BER of a Suzuki stack, at a given working frequency. So the resetting time and the BER are more important than the turn-on delay and the power consumption for an optimized stack. In order to reduce the resetting time, the junctions have to be shunted with some small resistors. There then comes the trade-off between the output level and the resetting time, or upper operating frequency.

Based on the above analysis, optimized parameters of a 2×16 JJ Suzuki stack are selected as follows. The output voltage should be larger than 40 mV in order to trigger sufficient current in the following interface circuit without adding too much gate capacitance by using a larger MOSFET. The bias margins for a BER of 1×10^{-6} should be at least $\pm 10\%$. And the resetting time and the turn-on delay time of the Suzuki stack should be minimized.

After optimization, such a Suzuki stack is possible based on the simulation results using a 2.5 kA/cm^2 Nb process. The junction shunt resistance is optimized to 20 Ω, and the simulation frequency is 5 GHz. The BER simulation results are shown in Fig. 3.11. The error-function fit lines indicate that the margin for a 10^{-9} error rate is $\pm 11\%$.

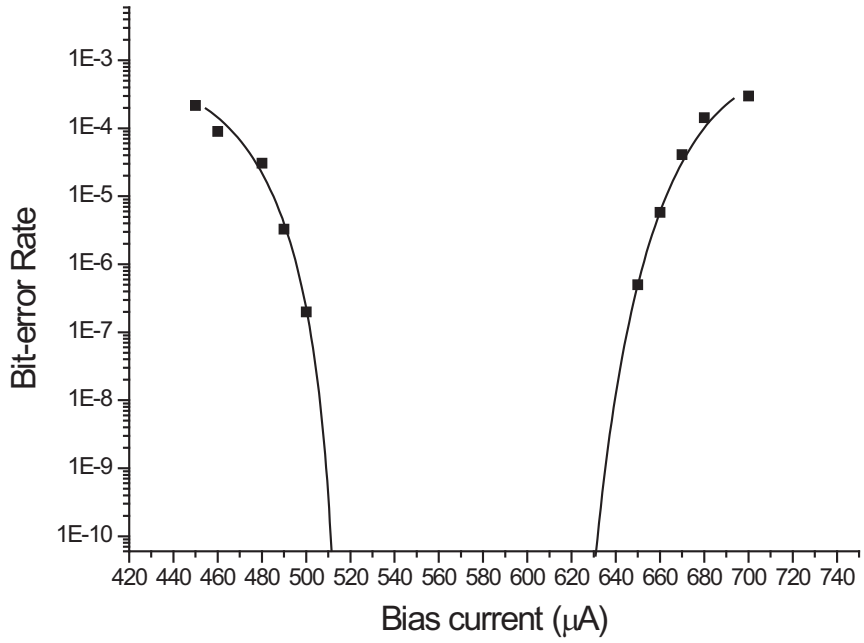


Figure 3.11: The BER simulation results of an optimized Suzuki stack using 2.5 kA/cm^2 Nb process working at 5 GHz. The junction shunt resistance is 20 Ω. The error function fit curves indicate that the margin for a 10^{-9} BER is about ±11%.

3.3 The second-stage amplifier

3.3.1 Candidates for fast amplification

Now that we have a synchronized 40 mV digital signal from the Suzuki stack, the job of the next part of the interface circuit is to finish the amplification, delivering a volt-level output. Given a 0.18 μm or more advanced CMOS process, the supply voltage could be as low as 1.5 V. With the scaling of CMOS technology and some special design as discussed in Chapter 2, a supply voltage lower than 1 V is possible. The voltage gain of this second stage amplifier should be about 25 with as small as possible delay time and power consumption.

Several candidates are available. The first obvious one is a traditional CMOS amplifier. For a CMOS amplifier, the voltage gain is not a problem; even the simplest one-stage common-source amplifier can achieve a voltage gain of 25. The problem is the bandwidth, the delay of the amplification, and the power consumption. The target for our hybrid memory system is 5 GHz operation with a sub-nanosecond access time. In order to operate a CMOS amplifier at 5 GHz with a voltage gain of 25, a more advanced process than 0.18 μm CMOS is required. Furthermore, since the delay time of CMOS address buffers and decoders will occupy a large portion of the total access time, there is little time left the delay of second-stage interface in order to keep the total access time in the deep sub-nanosecond region. The delay of a CMOS amplifier can be written as

$$t = \frac{Cv_{out}}{i_{out}} = \frac{Cv_{out}}{g_m v_{in}} = A_v \frac{C}{g_m} \tag{3.5}$$

which represents the total charge on the output node divided by total discharging current. In Eq. 3.5, C/g_m is a process-related parameter (inversely proportional to the cut-off frequency). Due to velocity saturation it is improved only a little by cooling to liquid-helium temperature. It is impossible for any CMOS process so far to implement an amplifier with voltage gain of 25 and delay time 100 ps. (The 100 ps delay is chosen because the delay of the CMOS decoder will be at least hundreds picoseconds). With the scaling of CMOS technology, such an amplifier may be possible some day, but the hybrid memory system simply could not afford its power (can be as high as hundreds of milliwatts).

Since the delay time is inversely proportional to the cut-off frequency of a process, why not use a process that has a high cut-off frequency; say, a SiGe HBT process or even a GaAs process? The answer is excessive power dissipation. It is true that an HBT device has much higher cut-off frequency and studies has been made to verify that HBT circuits work even better at 4 K. [51] But in order to build an amplifier with less than 100 ps delay and a voltage gain of 25, a fairly large amount of power has to be consumed. Actually, there are commercially available amplifiers like HP/Agilent 83017A (0.5-26.5 GHz 25 dB gain), but they are not on-chip because they comsume so much power (9 W) that you cannot afford to build them on chip. [52]

Another candidate would be a superconducting circuit. One candidate would be

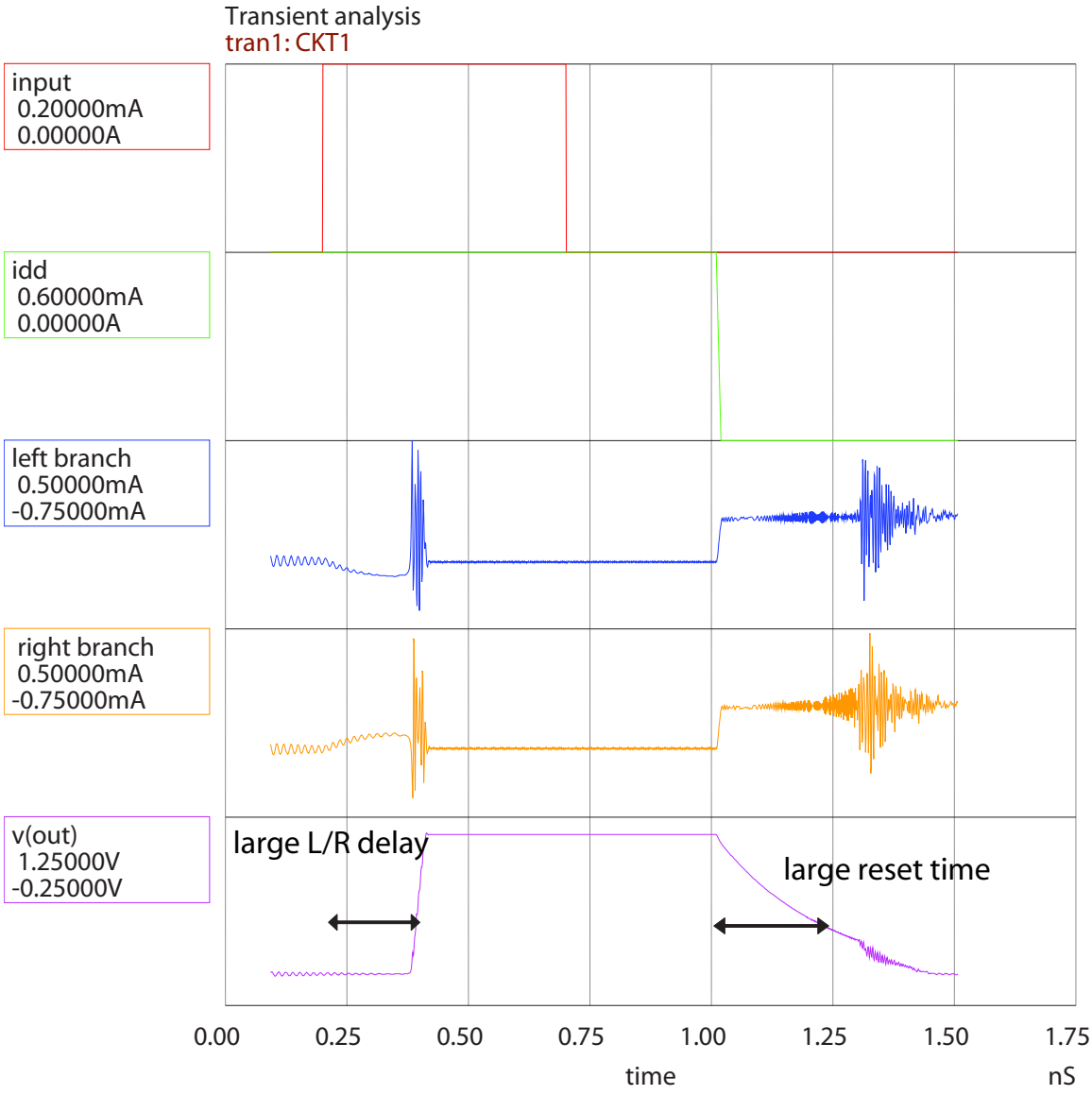


Figure 3.12: The simulation result of a 2×400 JJ Susuzki stack. Both the turn-on delay time and the resetting time increase dramatically due to the larger inductance and effective resistance, which prevents it from being a candidate for the interface amplifier.

a long Suzuki stack. A 2×16 JJ Suzuki stack gives about 40 mV output, and we could consider using a 2×400 JJ Suzuki stack which would give a higher output level. However, it would introduce more delay time, resetting time and maybe more important, greater possibility of punchthrough and, therefore, more errors. Fig. 3.12 shows the simulation of a 2×400 JJ Suzuki stack that gives about 1.2 V output with a much longer delay time and resetting time than those of the 2×16 JJ Suzuki stack. This simulation assumes a 10 fF parasitic capacitance at the output node. The delay time, Eq. 3.2, cannot ignore the L/R current redistribution delay, due to the large total inductance in an 800 Josephson junction loop. As is shown in Fig.3.12, the redistribution time dominates the total delay while the charging delay is only a small fraction of the total delay. The resetting time, on the other hand, suffers even more when the stack is higher because of the unavoidable parasitic capacitance at the output node. In Eq. 3.2, the contribution by the parasitic capacitance is the dominant part, giving a resetting time as large as 0.5 ns, preventing the Suzuki stack from being operated at 5 GHz. Even if the circuit is operated at a lower frequency, say 1 GHz, the large resetting time enhances the possibility of punchthrough and, therefore, increases bit error rate of the circuit.

The candidates discussed above cannot complete the amplification quickly with affordable power consumed. Some other circuit must be designed in order to complete the interface circuit without compromising too much performance or consuming too much power. In order to design such a circuit, it is helpful to summarize why the

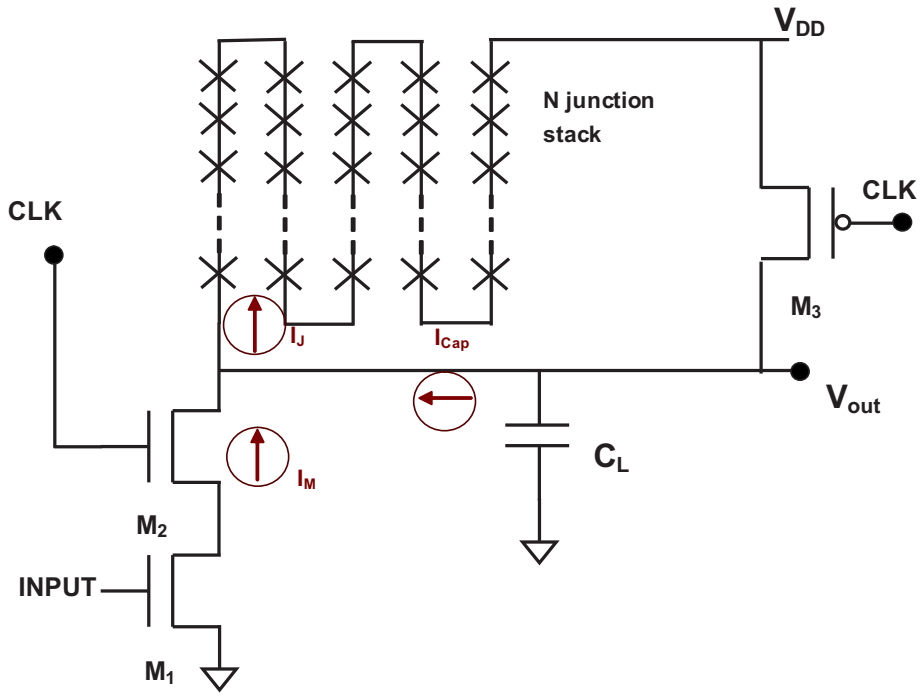


Figure 3.13: The schematic of a hybrid second-stage amplifier, which amplifies a 40 mV input to a volt-level output voltage. It is an inverting amplifier. The C_L represents all parasitic capacitance and the gate capacitance of the following circuit. The precision of the amplifier is not a problem because the following CMOS digital circuits have fairly large margins. In the current design, the long junction array has 400 junctions which gives a 1.2 V output. The dynamic currents during a switching process are shown on the schematic, as described in the text.

other candidates fail. The semiconductor amplifier has an intrinsic gain-bandwidth-product limit that is process related and there is nothing we can do in design. For the long Suzuki stack idea, the main problem is the L/R current redistribution time and the resetting time. If some new Josephson-junction-based design is used as the second-stage amplifier, these two problems must be overcome. First of all, L/R delay must be minimized, either the loop inductance is lower or the resistance is increased. Secondly, something with smaller resistance must be shunted with the junction array in order to reduce the resetting time. Our so-called hybrid amplifier solves the two problems. Fig. 3.13 shows the schematic of such a hybrid design, which was originally proposed by U. Ghoshal. [53]

The operation of this hybrid amplifier is as follows: When the clock is high, M_1 is biased so that the current in M_1, M_2, and the N-junction array is a little less than the critical current of the N junctions, typically $0.8I_c$, as in previous discussion for Suzuki stack. When the 40 mV input arrives at the gate of M_1, this NMOS transistor acts like a voltage-controlled current source, increasing the current by $\Delta i_{M1} = v_{in} g_m$, where g_m is the transconductance of M_1 at a given bias point. Note that Δi_{M1} must be much greater than the process variation of critical currents in order to achieve a robust amplifier, and should provide enough overdrive to switch the output stack at high speed. When the current in the junctions exceeds the critical current, all of the junctions switch and the output voltage is lowered by NV_g. Of course, this is still a latching circuit without the ability to reset itself, but the M_3 PMOS transistor is

provided to speed the resetting. When the clock is low, M_2 is off and M_3 is on, the output voltage will be charged up to V_{DD} via M_3 and pull all junctions back to the superconductive states. When the clock is high again, all the junctions are biased at $0.8I_c$ and are waiting for the next input.

The transistor M_1 acts as a current source with a large source resistance (the output resistance of M_1 depends on the channel length and is on the order of kilohms) therefore effectively reducing the L/R delay time. By adding one more NMOS (M_2) making a cascode structure, the source resistance of the cascode structure is enhanced by a factor of $g_m r_o$, helping to decrease the L/R delay even more, and more importantly, minimize the so-called Miller effect. Without the cascode structure, because of the inverting amplification between the gate and the drain of M_1, the gate-drain capacitance will be amplified by the Miller effect to a larger capacitance, therefore, degrade the performance of the preceding Suzuki stack.

The PMOS M_3 is an important device in this circuit because it greatly reduces the effective discharging resistance from NR_{sub} to $NR_{sub}||R_{PMOS}$, without adding too much load capacitance since the drain/source junction capacitance is frozen to almost zero at 4 K, thus decreasing substantially the resetting time and the punchthrough possibility.

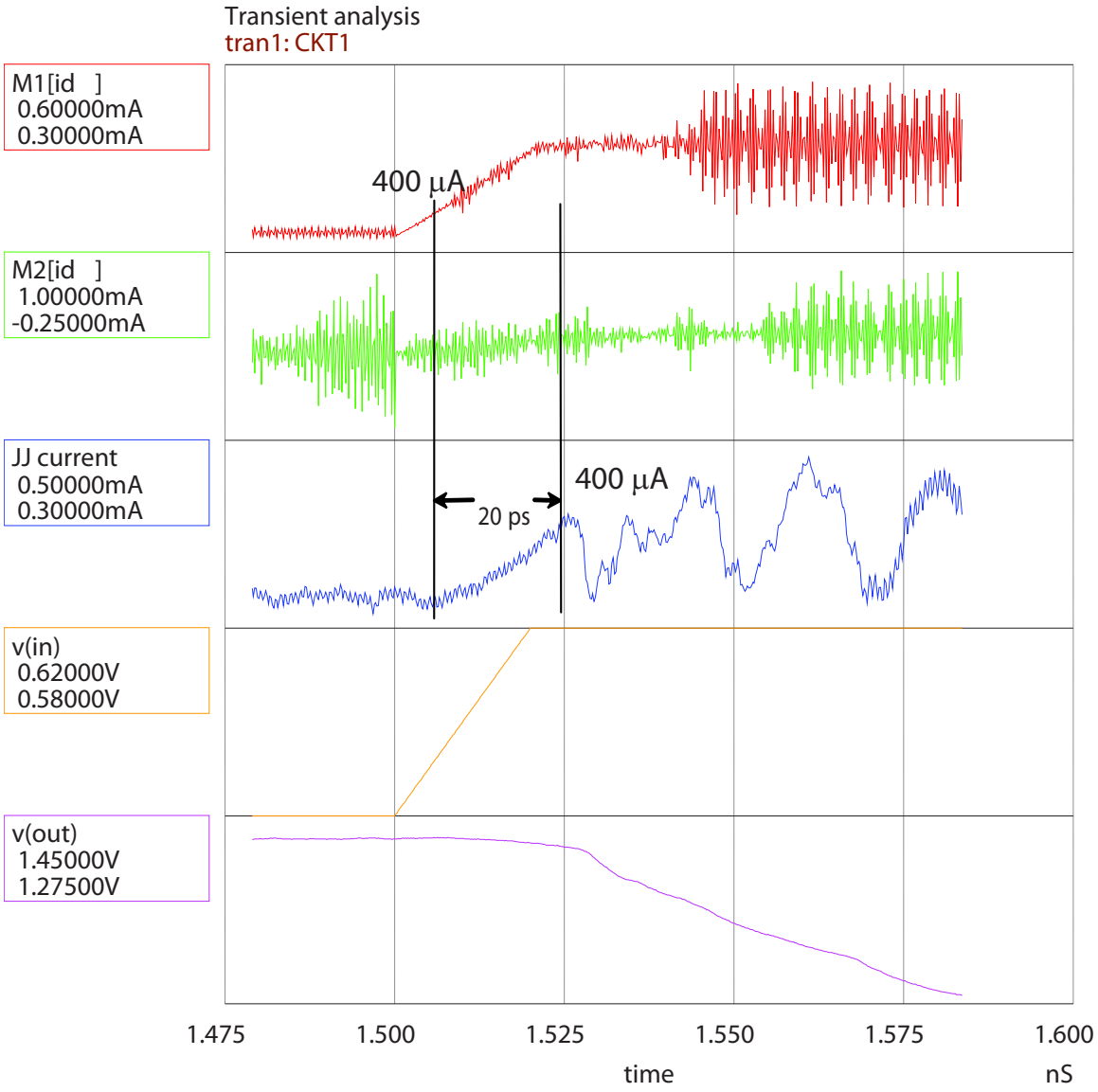


Figure 3.14: The simulation results of a 400-JJ hybrid amplifier. The current redistribution time depends on the drain-gate capacitance of the two N-type transistors; the 20 ps delay from when current in M_1 is 400 μA to when the current in junctions is 400 μA is caused by the parasitic capacitance of the transistors, not the inductance.

3.3.2 Delay of a hybrid amplifier

The delay analysis of this hybrid second-stage amplifier is almost the same as the delay analysis of a Suzuki stack. The first phase will be the L/R delay, since the cascode output impedance is large, this current redistribution delay is very small. A typical value is about 1 ps and can be ignored. However, the cascode structure introduces some delay due to the gate-drain capacitance of M_1 and the load capacitance. The reason is that this capacitance, although small, introduces a current that flows into the drain of M_1 therefore adding a delay between I_{M1} and I_{M2}. The load capacitance causes another delay between I_{M2} and I_J. These two small delay times, according to the simulation results, totals only about 20 ps, shown in Fig. 3.14. After the switching of the N-junction stack, the output node voltage drops by NV_g in a short time. This short delay depends on how fast the capacitor is discharged.

Applying KCL at the output node, the current equation can be written as

$$I_{cap} = I_J + I_{M2} \tag{3.6}$$

where I_{cap} is the total discharging current only during the discharging process, I_J is the dynamic current that flows upward to V_{DD} node, and I_M is the dynamic current that flows from ground to output, shown in Fig. 3.13. Since the voltage is decreasing, the current in the cascode transistors must decrease somewhat, according to the I-V characteristics of a cascode transistor structure. That is the reason the dynamic current is flowing upward. So part of the discharging current goes to the 400-junction

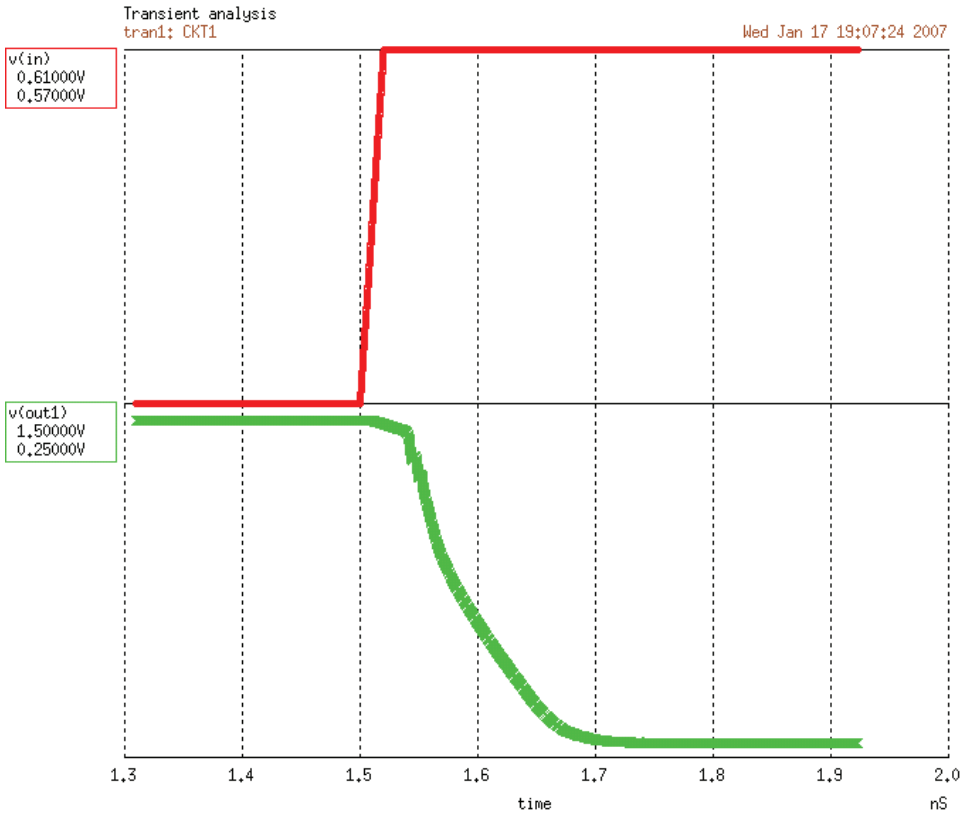


Figure 3.15: The simulation results for a 400-JJ hybrid amplifier. The total delay is about 70 ps for a load capacitance of 20 fF.

stack, causing a current decrease in the junctions. But the junctions could not hold to the voltage state if the current inside were almost zero. Based on the first-order analysis above, we conclude that the discharging current is some fraction (k) of the critical current, so the discharging time is

$$t_{amp} = \frac{NV_g(C_J/N + C_L)}{kI_c} = t_J/k + \frac{NV_gC_L}{kI_c} \tag{3.7}$$

where t_J is the delay time for switching a single junction, as shown in Eq. 1.3, and is a process-related parameter, about 6 ps for a 2.5 kA/cm^2 Nb process. And k is a constant less than, but close to 1, according to simulation results. Fig. 3.15 shows the simulation results for a 400 JJ stack with a 2.5 kA/cm^2 Nb process, assuming a 20 fF load capacitance.

Fig. 3.16 shows the delay versus load capacitance, the straight line confirms that the delay is linearly proportional to capacitance, as indicated in the above analysis.

Two conclusions can be drawn from Eq. 3.7. First, for large-number junction stacks, the delay time is dominated by the parasitic capacitance discharging time . For a 400-junction array with critical current 400 μA, the parasitic capacitance cannot be larger than 50 fF in order to have a delay time less than 100 ps. Secondly, in order to have less delay, a larger critical current and smaller parasitic capacitance are required.

Resetting time, on the other hand, is totally controlled by the effective resistance of the shunt PMOS device. The total sub-gap resistance for a 400-junction stack

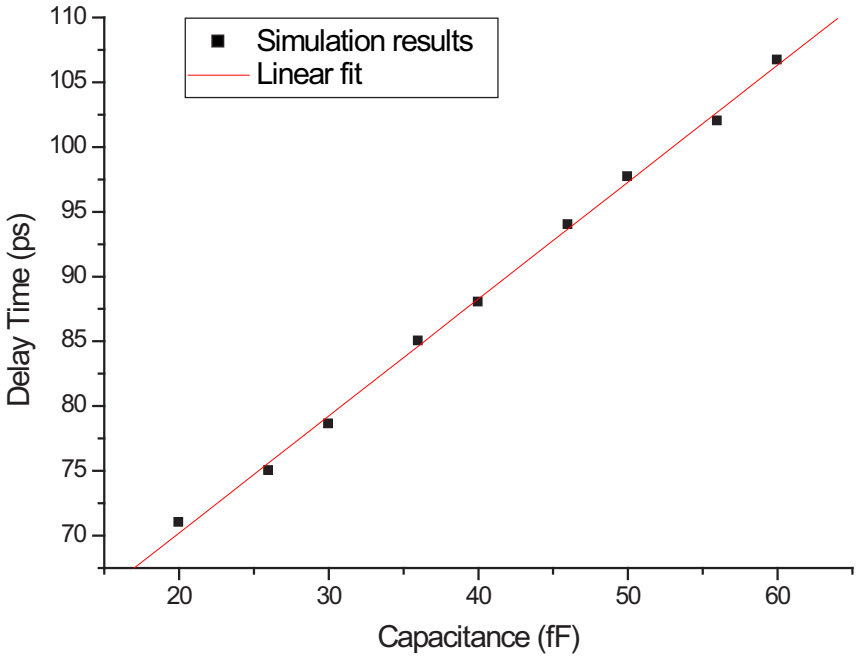


Figure 3.16: The relationship between the delay time and the parasitic capacitance of a 400-JJ hybrid amplifier.

is about 120 kΩ given that the sub-gap resistance for a single junction is 300 Ω (a typical number for a 2.5 kA/cm^2 Nb process). If the PMOS were not there, even a small parasitic capacitance like 10 fF will cause a 1.2 ns resetting time. However, by adding a shunt PMOS and making the size large enough that the effective resistance of the PMOS is only on the order of 1 $k\Omega$ or less, the resetting time can be suppressed to an acceptable level. Because the drain/source capacitance of MOSFETs decreases to almost zero at 4 K, adding this large PMOS does not contribute too much to the load capacitance.

3.3.3 Power consumption

Just as for the Suzuki stack, the power consumption of this hybrid amplifier depends heavily on the quasi-static power, which is the product of bias current (a fraction of the critical current) and the supply voltage. There is a trade-off between decreasing the delay time and decreasing the power consumption. The design of the hybrid amplifier will have to make a trade-off between these two aspects, depending on the system requirements. However, one can take advantage of the sharp subthreshold characteristic of 4 K CMOS to bias the circuit at just below threshold voltage so that the bias current is almost zero, therefore reducing the static power. This method, however, has two disadvantages. First of all, it requires a larger M_1 because the g_m is smaller at lower bias and larger M_1 increases the gate capacitance that eventually increases the delay and resetting time of the preceding Suzuki stack. The other

shortcoming is the extremely low bias margin. Since the bias has to be slightly (10 mV) under the threshold voltage so that the 40 mV input will turn the transistor on and deliver large enough current to switch all the junctions, the margin is something like 550 $mV \pm 10\ mV$, only $\pm$ 2% margin. A precise and self-biased circuit must be designed in order to control the bias in the right margin. And this bias scheme has to tolerate the threshold voltage local variations of CMOS chips, which is another important practical issue.

3.3.4 Margins

The low-frequency operating margins of this hybrid amplifier can be analyzed similar to the margins of the Suzuki stack, in terms of bias current through the 400-JJ array. Here, the bias current is controlled by the dc bias voltage on the gate of M_1. And 40 mV should deliver large enough drain current to switch all junctions, so the bias margin is less than 40 mV, on a bias voltage of about 0.7 V. This is only a 6% margin, which is lower than that of a Suzuki stack or any other superconducting circuits. This is an inherent problem with the hybrid amplifier, as long as the threshold voltage is as high as about 0.6 V. If a special CMOS process were available and were designed for 4 K operation, the threshold voltage could be as low as 50 mV, much larger percentage margins would be possible. For current commercial CMOS chips working at 4 K, this small margin problem for the hybrid amplifier is unavoidable.

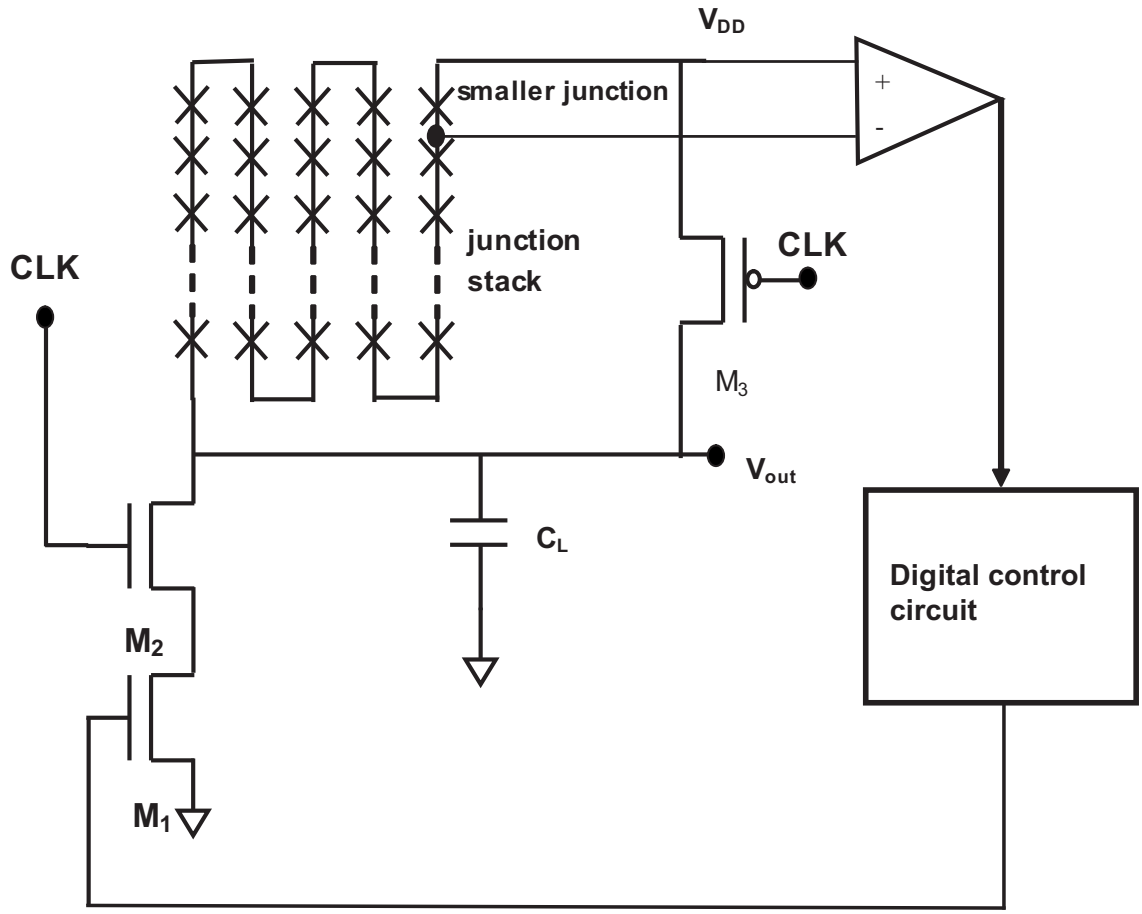


Figure 3.17: The self-bias scheme to precisely control the dc bias point of a 400 JJ hybrid amplifier, in order to solve the small-margin problem.

Since the margins are really small, a good bias scheme is necessary to ensure the right bias. A self-bias scheme shown in Fig. 3.17 is proposed. In the junction stack, the top junction has a critical current of 0.8 I_c and others have critical current of I_c. The control box will keep increasing the bias voltage at each clock cycle until the bias current is greater than 0.8 I_c and switch the top junction, an 2.8 mV output voltage will be amplified by a traditional CMOS amplifier and fed to the control circuit to prevent further increase of the voltage by a digital logic circuit. [55] After it settles down, the bias voltage at the gate of M_1 is the desired value and the bias current is 0.8 I_c as designed. The entire amplifier and control circuit can be easily implemented in commercial CMOS.

3.3.5 Clock feedthrough

Another design issue for a hybrid amplifier is the clock feedthrough. Clock feedthrough is a common phenomenon in CMOS circuits, especially in switched capacitor circuits. Due to the presence of the gate-drain capacitance, although small, the output voltage suffers an error voltage of $\Delta V = \frac{C_{gd}}{Cgd+C_L} V_{CLK}$ at clock rising or falling edge, and the falling edge is more important due to the channel charge redistribution. This error voltage contaminates the accuracy of the sampling and, therefore, compromises the precision of the sampling.

For our hybrid amplifier, however, the precision is not a big problem. we do not need precise amplification for interfacing between SFQ to CMOS digital signals.

Instead, any voltage that is larger than 1 V can drive 0.18 μm CMOS digital circuits work because of the large margins of CMOS digital circuits. And smaller voltage will be amplified to full V_{DD} swing by the following inverters. For the hybrid amplifier, the main problem caused by the clock feed-through is the JJ switching associated with it. For a clock rising or falling event, the clock signal applied to the gate of M_2 sees a small capacitance C_{gd} in series with a large capacitance C_L. The capacitive divider forces a small voltage step on the load capacitor. The charge on the capacitor is not a problem in Josephson junction circuits. Rather, the problem is how these charges find a way to ground, and the current associated with that.

The charge on the output node see three paths to ground, the PMOS, the cascode NMOS transistors, and the long junction array. The MOSFET path has a large resistance (on the order of kilohms); on the other hand, the junction array only has a sub-nanohenry inductance and zero resistance before junctions switch. Therefore, the charges will flow preferentially through the junction array to the V_{DD} node (effectively ground for time-dependent signals). Let's call this current a negative current. The total current injected into the gate-drain capacitance depends on the voltage step and its rise time. If this injection current is larger than the critical current, the junctions switch, but with an opposite direction, resulting a voltage increase. In other words, the presence of the capacitive divider forces the junctions to switch to maintain the error voltage, if the charge flows to ground quickly enough. Fortunately, this situation does not last long. After the clock increases to a value such that the current in the

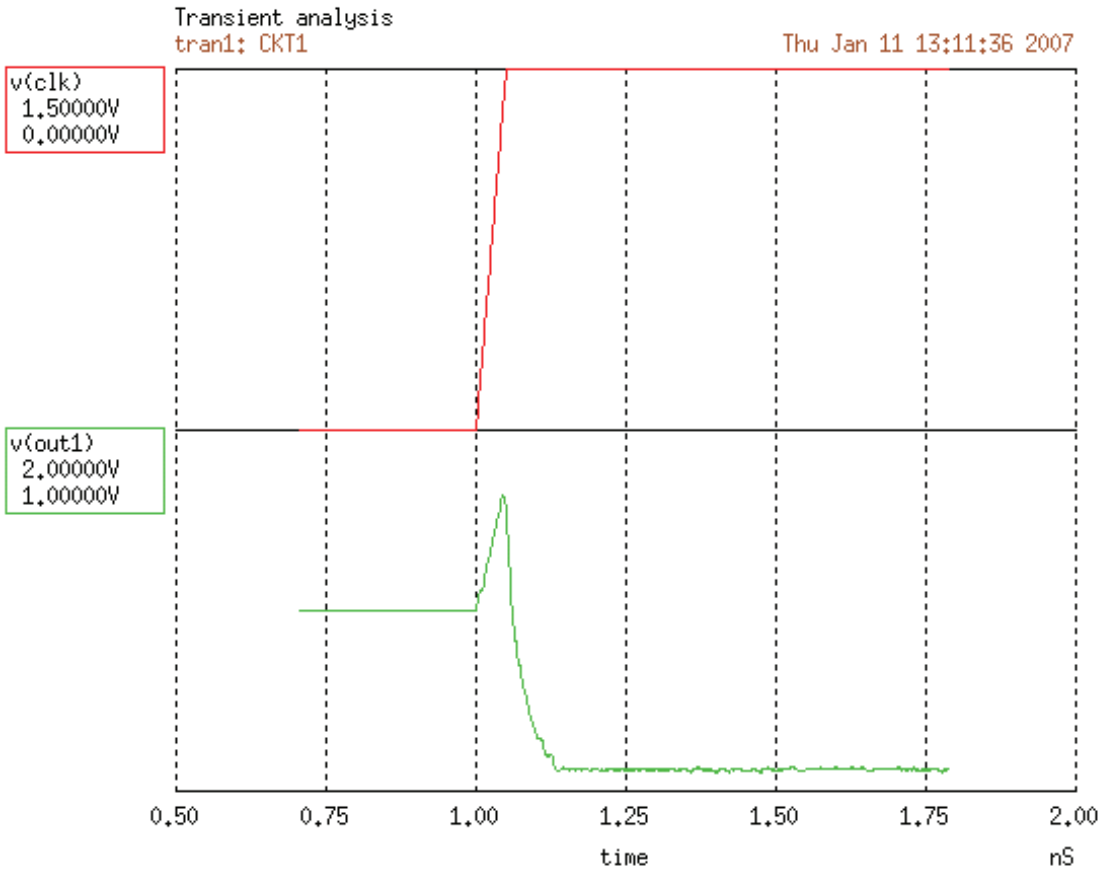


Figure 3.18: The simulation shows a clock-feedthrough induced output drop if the rise time of the clock is too small.

NMOS transistors is positive and the current in the PMOS is zero, the junctions will have a positive current flow. However, due to the transmission-line delay, at a given time, currents in the 400 junctions are not all identical. After the positive current $+0.8I_c$ arrives, it is possible that some junctions will remains superconducting with a bias current $+0.8I_c$ while others switch to the voltage state, resulting a voltage drop at the output node. This is the most important problem caused by the clock feedthrough. Fig. 3.18 shows the output drop for a clock with a 50 ps rise time. For larger rise time, say 500 ps, there is no such a voltage drop, because the current injected into the junction array is not large enough to switch the JJs.

The mechanism of this effect can be explained well in an intuitive model using pendulum analog. In the analog, the junction array looks like a pendulum array connected in series with springs. When a clock step event occurs, a negative torque(representing the injection current) is applied to the pendulum array, causing all pendulums to rotate clockwise until the phase exceeds 90 degree. Due to the delay caused by the spring, some of the pendulums keep clockwise rolling and others roll backward to form oscillations. All the pendulums keep rolling or oscillating until the bias current controlled by the clock is increased (representing a positive torque is applied to all the junctions). This applied torque forces the rolling and oscillating pendulums to slow down and eventually settles junctions down to the appropriate angle ($sin^{-1}(0.8)$). However, the presence of the springs delays the propagation of the rolling and oscillating, as well as the applied torque. At a given time, every pendulum has a different

position, torque, and velocity. For some pendulums, the applied torque arrives at such a time that it pushes the pendulum to the right position with just a little oscillation. But for some other pendulums, it is possible that the applied torque is applied too fast or too hard such that the torque not only slows down the clockwise rolling, but also forces the pendulum to roll counter-clockwise, which represents that the junction is switched to the opposite direction, giving a lower voltage at the output node.

Fig. 3.19 shows the simulation results of phases of different junctions. It shows that some of the junctions settle down to the right position while some others switch to the voltage state after the positive bias current appears. The sine values of the junction phases show the switchings of junctions. The first junction is switched by the injection current and results in a negative voltage drop, and the second junction is also switched by the injection current, but keeps oscillating around 53 degree, which does not contribute to the voltage drop.

The mechanism of the clock feed-through is qualitatively clear, but it is hard to build a simple model to quantitatively simulate the whole process except by a numerical simulation like WRSPICE. However, in order to minimize clock feedthrough, the qualitative analysis gives us some basic idea. First of all, the gate-drain capacitance is the most important parameter in this effect. To suppress the clock feedthrough, a smaller gate-drain capacitance is preferred. Note that both the cascode NMOS and the resetting PMOS contributes to this capacitance. We need smaller NMOS and PMOS devices. But resetting time requires a larger PMOS. Secondly, a larger clock

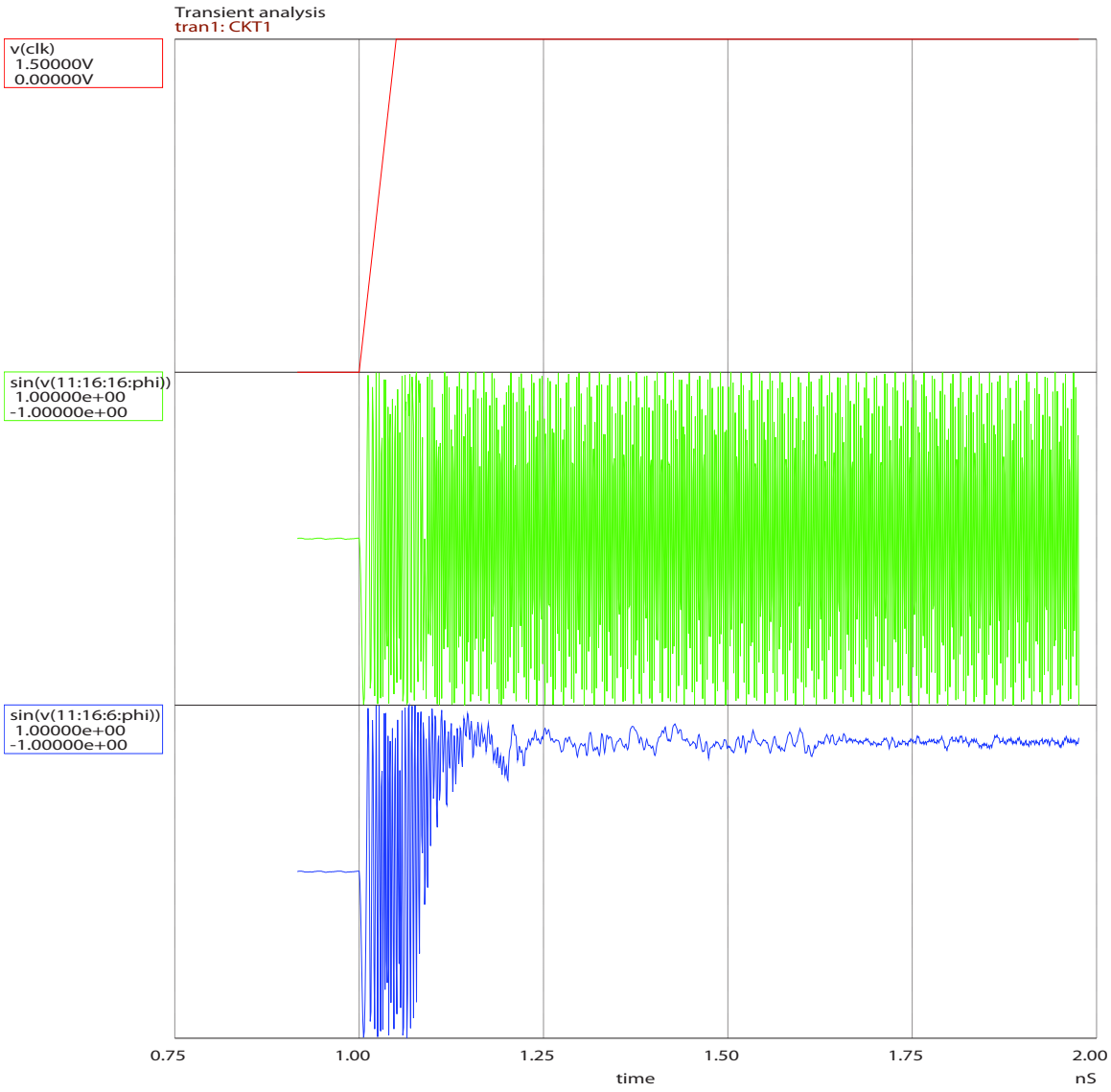


Figure 3.19: The simulation shows a clock-feedthrough scene, where some junctions suffer from inverse switchings. The clock arrives at 1 ns with a rise time 50 ps. When the clock reaches its full value, some junction, represented by the second line, remains the voltage state and some junctions, represented by the third line, do not.

rise time is preferred to reduce the displacement current in order not to switch the junctions oppositely. This is apparently not an option for high-frequency operation. Those two solutions are intended to reduce the injection current from the clock signal. Since there are trade-offs that we cannot avoid, we just cannot suppress the injection current too much. In other words, we have to live with the fact that there is always some injection current that might cause some random behavior in the junction array.

There is another way to suppress the feedthrough, which is to increase the critical current. Increasing the critical current makes the injection current relatively smaller and the possibility of the injection current switching is lowered. However, as we discussed before, though increasing the critical current reduces the delay time, it increases the power consumption. As will be discussed later, the interface circuit power is the main part of the total power of a hybrid system. Therefore, to increase the critical current is not a good idea in terms of power.

Now we have a limit in decreasing the injection current, we also have a limit in increasing the critical current, it looks like there will be some random behavior anyway in the junction array, after the clock rises. Fortunately, although we cannot avoid the random behavior in the array, at least we have a way to reduce it. The way is to add a shunt resistor in parallel with each junction, just as is done to suppress punththrough. In the pendulum analog, to add a shunt resistor is to increase the damping. Smaller resistance corresponds to larger damping. When a torque (corresponding to an injection current) is applied to a pendulum with larger damping,

the oscillation will settle down quicker because every oscillation consumes more energy than with less damping (the Q factor drops). If oscillations settle down in a shorter period, an opposite applied torque has a smaller possibility to cause an opposite oscillation, representing a voltage drop. However, we cannot shunt with too small resistors, otherwise the output voltage will drop accordingly.

Although the clock feedthrough causes random behavior inside the junction array and the voltage drop before the input arrives, the switching of the second-stage amplifier remains unharmed. After careful design, the voltage drop caused by the clock feedthrough can be as small as 0.1 V, which is not a problem for the switching of the following amplifier.

3.4 Fighting with parasitics

3.4.1 Parasitic calculations

As was discussed in previous sections, both parasitic capacitance and inductance play important roles in the Suzuki stack and in the second-stage amplifier. It is, therefore, very important to study how to calculate those parasitics as well as how to change them.

In a standard Nb process, there is a niobium layer as a ground plane called $M0$ layer underneath all superconducting circuits on chips. If we keep it that way, every junction will have a parasitic parallel-plate capacitance to ground associated with it,

and the value can be written as

$$C_g = A_J \epsilon_{SiO2} / t_{SiO2} \tag{3.8}$$

Given a 2.5 kA/cm^2 process, every 0.1 mA critical current junction has a capacitance to ground of 5 fF; The parasitic inductance can be written as

$$L_p = \frac{\mu_0 d}{w}(1 + \frac{2\lambda}{d}) \tag{3.9}$$

where λ is the penetration depth of the superconducting material and d is the distance between two superconducting metal layers. (This equation is only precise when the thickness of the metal layer is much larger than the penetration depth.) For a 2.5 kA/cm^2 Nb process, the inductance is around 0.7 pH per junction.

The most convenient way to reduce capacitance is to remove the ground plane underneath the junctions. The capacitance to ground is then reduced due to the much longer flux path. And the junction array can be treated approximately as a co-planar waveguide, because the gap between the junction array and the edge of the ground plane is much larger than the vertical distance between the junction array and the ground plane, as shown in Fig. 3.20. By removing the ground plane and modeling the array as a co-planar waveguide, the capacitance and the inductance per unit length can be written as [54]

$$C_g = 2\epsilon_0(\epsilon_r - 1)\frac{K(k)}{K(\sqrt{1-k^2})} \tag{3.10}$$

$$k = \frac{\sinh(\frac{\pi w}{4h})}{\sinh(\frac{2\pi(w+2s)}{4h})} \tag{3.11}$$

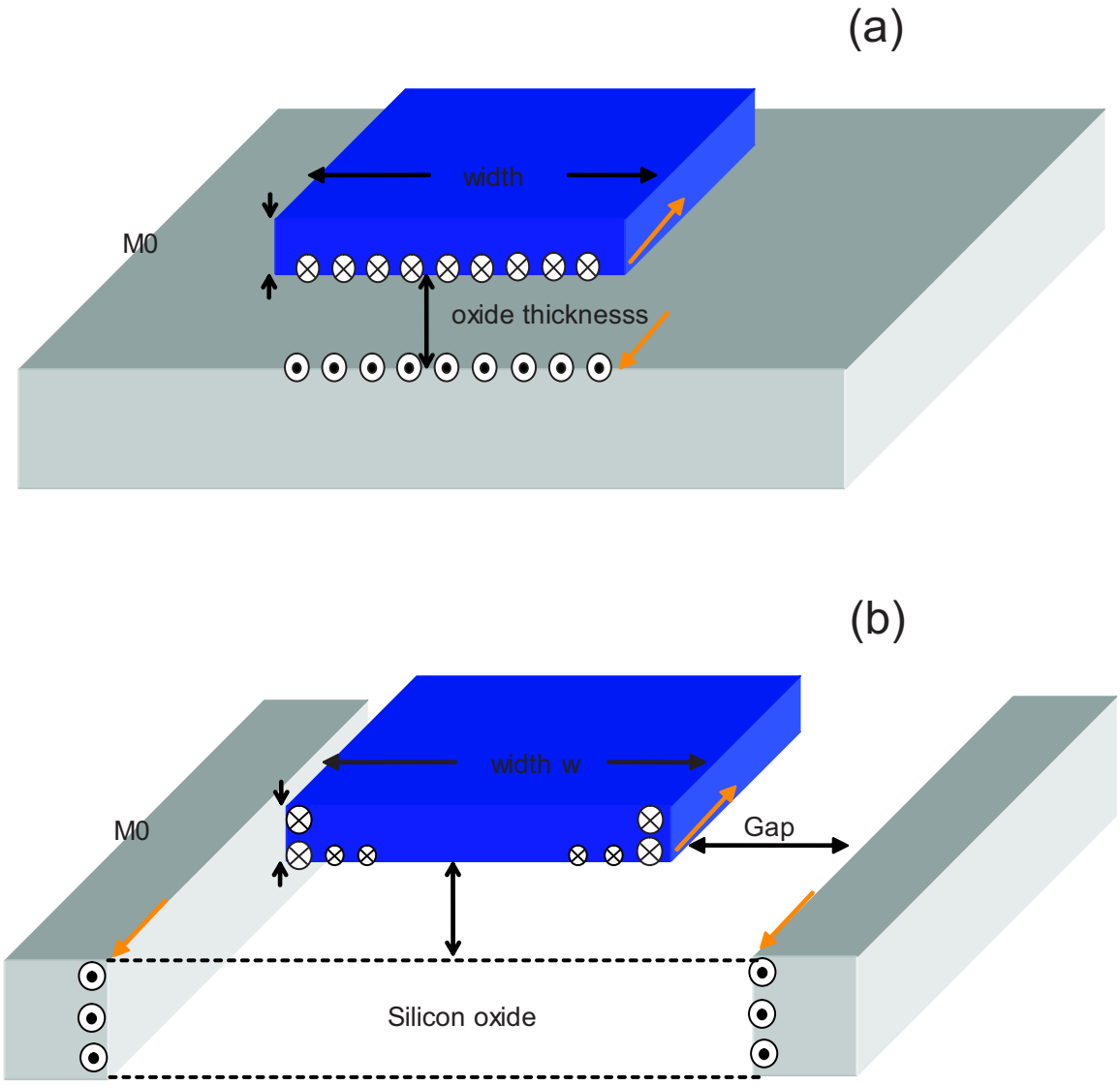


Figure 3.20: The structure of a) junction array with ground plane underneath and b) junction array with the underneath ground plane removed, leaving a gap between the array and the ground. Charges and currents are located roughly as shown. The picture is not to scale. The gap is actually much larger than the oxide thickness.

$$L_p = Z_L^2 C_g \tag{3.12}$$

$$Z_L = \frac{\eta_0 K(k')}{4K(\sqrt{1-k'^2})\sqrt{\epsilon_r}} \tag{3.13}$$

$$k' = \sqrt{1-(\frac{w}{w+2s})^2}, \tag{3.14}$$

where w is the width of the junction array, s is the gap width between the edge of the ground plane and the edge of the junction array, K function is the elliptical integral of the first kind. Although it is a simplified model, it gives a clear physical picture and fits numerical simulation results within 20%.

As seen in Eq. 3.10, the capacitance decreases while the inductance increases when the gap increases, as shown in Fig. 3.21. As discussed previously, the smaller capacitance helps to reduce the delay and resetting time for the interface circuit. The larger inductance, however, does not degrade the performance of the interface circuit very much. For the Suzuki stack, it is the loop inductance that really matters in terms of delay. The current redistribution time depends on the loop inductance, as shown in Eq. 3.2. The parasitic inductance calculated above is the inductance of the junction array. The loop inductance depends on the geometric shape of the array and does not change whether or not there is a ground plane underneath. While the inductance has some effect on the delay of the Suzuki stack, this delay is a small part of the total delay. For the second-stage amplifier, the junction array inductance has little effect on the delay time due to the large output impedance of the driving cascode structure. Simulations also verify that even 10 times larger inductance does not increase the delay time by 1%.

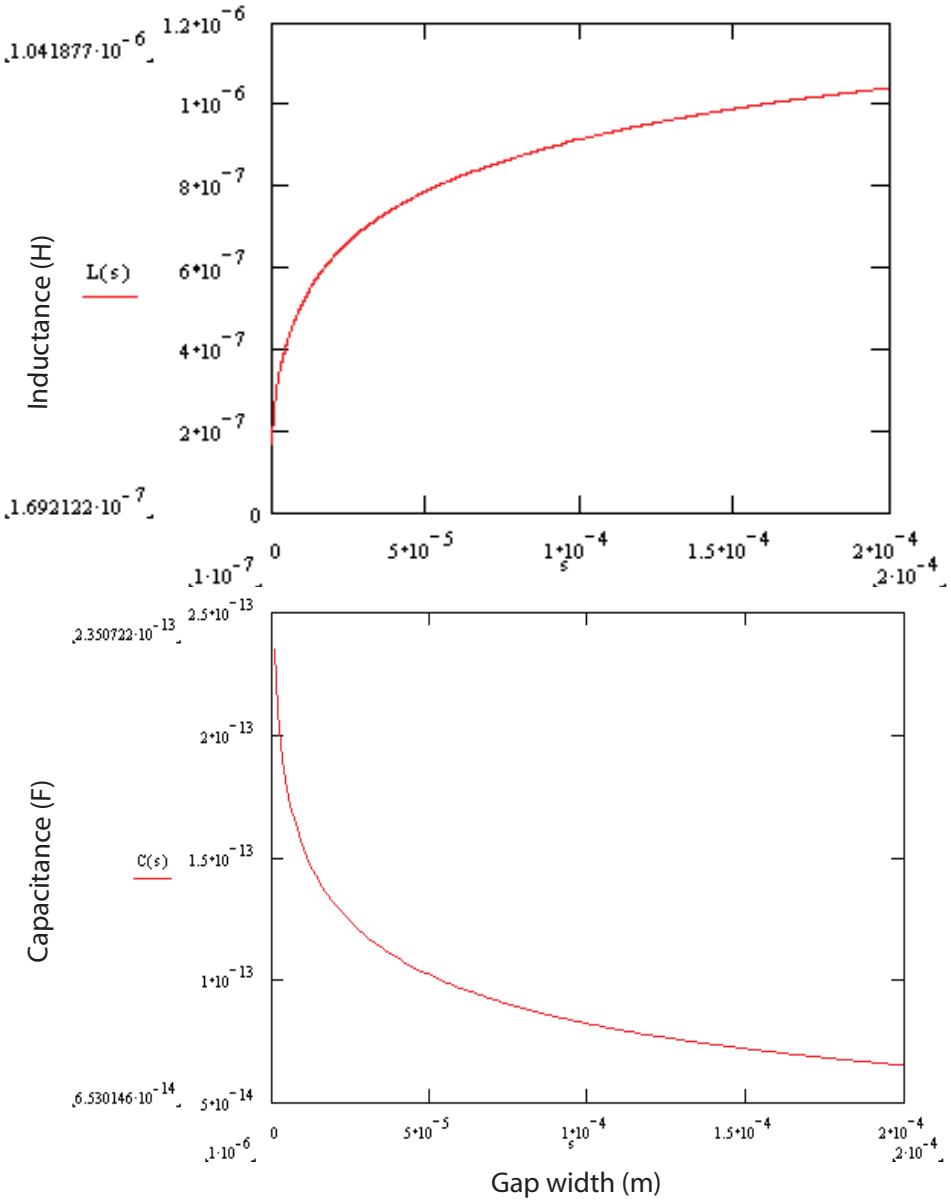


Figure 3.21: The calculated inductance and capacitance of the 400-junction array with the ground plane removed as in Fig. 3.20(b). [54]

Therefore, the speed performance of the interface circuit, including the Suzuki stack and the second-stage amplifier, benefits from smaller capacitances, while the increase of inductance has little negative effects on the speed performance. This makes this "removing-ground" approach very promising to speed up the interface circuit.

3.4.2 Layout/process techniques to minimize parasitics

Removing the ground plane underneath desired circuits can be done in the layout process. There is something more we can do during the fabrication process to reduce the capacitance even more. One way is to make a trench around the junction array by etching the silicon substrate. The electric flux finds a path that requires minimal energy. And since the dielectric constant of air is 12 times less than that of silicon, the electric flux tend to travel through silicon. By adding a trench, the effective gap between the ground-plane edge and the array edge is increased. Assuming the flux in air does not change much, the total capacitance will be decreased by a factor of 50% for a 50 μm deep trench with a 50 μm gap between the array and the ground plane.

One more trick can be played in order to reduce capacitance. After the wafer is finished with all the circuit deposition process, one last process can be added. By using KOH etching from the back of the wafer and precisely controlling the etching velocity, a 1 μm-thick membrane can be fabricated as shown in Fig. 3.22. By removing the silicon under the array, the effective dielectric constant is reduced approximately by a factor of two, so will be the total capacitance.

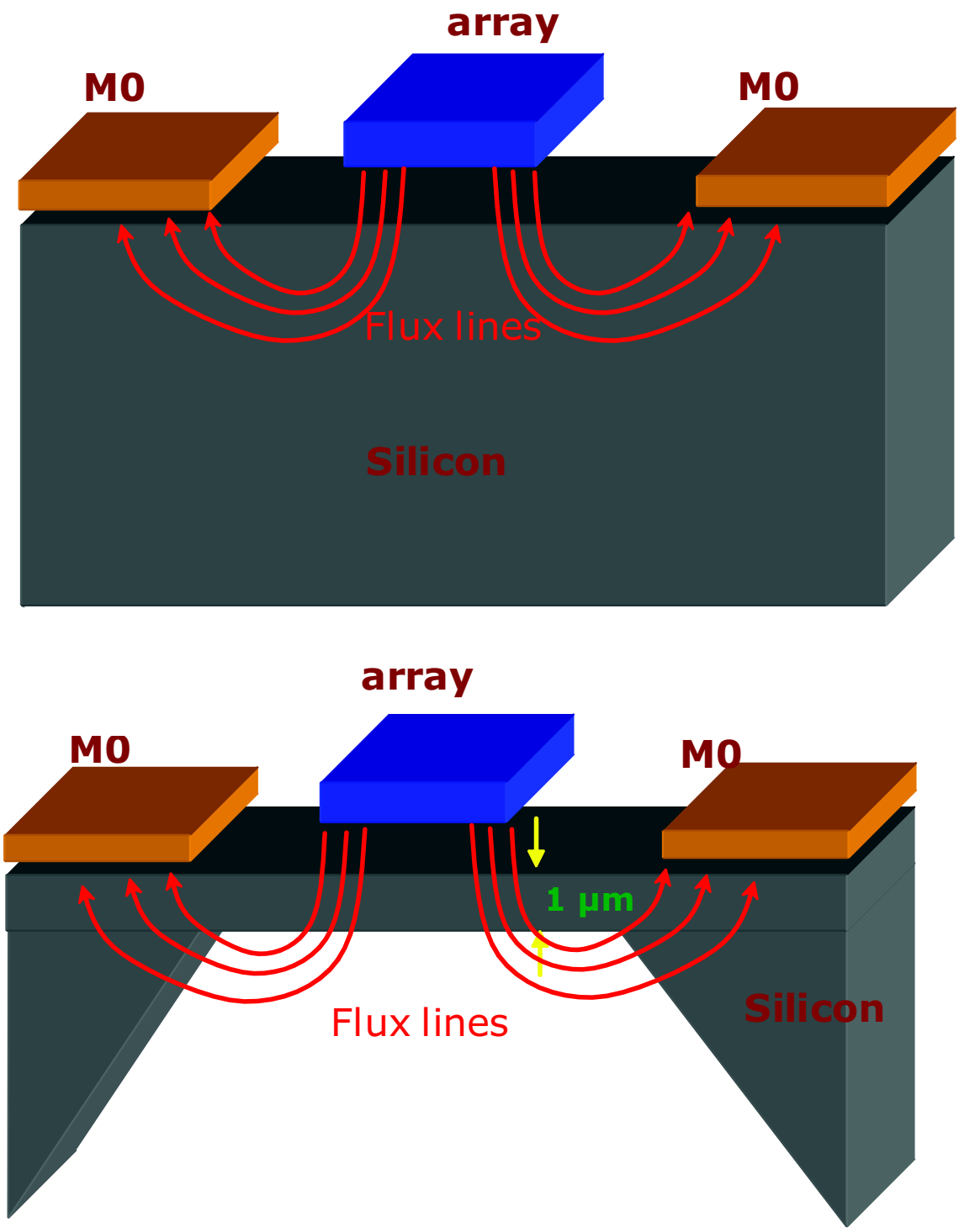


Figure 3.22: The normal and membrane substrate structure for a junction array. The picture is not to scale, the thickness of the membrane is much smaller than the gap. The flux lines show that the capacitance of the membrane structure is smaller than the normal one.

3.4.3 How to represent parasitics in simulations

As was discussed before, with longer junction stacks, the parasitic capacitance to ground and inductance are more important because the total effective resistance for charging and discharging is larger. How to represent parasitic parameters in simulations is an important question. One way is to represent the total capacitance to ground associated with the output node and the total parasitic inductance in series with the junction array simply as lumped elements at the end of the array, as been done in previous analysis and simulation. This method provides a simple model for the first-order analysis and gives a quick physical picture of the circuit. However, it is not as accurate as the distributed model, in which the total capacitance and inductance are calculated and distributed among the junctions.

The distributed model for parasitic calculations better represents the real situation. In this distributed model, the distributed inductance and capacitance comprise a transmission line with a characteristic impedance $\sqrt{L/C}$. Consequently, a matching resistor of value $\sqrt{L/C}$ should be added between the V_{DD} node and the top of the junction array in order to minimize any possible reflection. In our current layouts, $\sqrt{L/C} = 60\ \Omega$ so the matching resistor causes only a supply voltage drop of about 24 mV.

The discharge time in the first-order analysis should be smaller for a distributed model because each junction sees a different voltage drop. The worst contributor to the delay time is the capacitance associated with the junction closest to the output

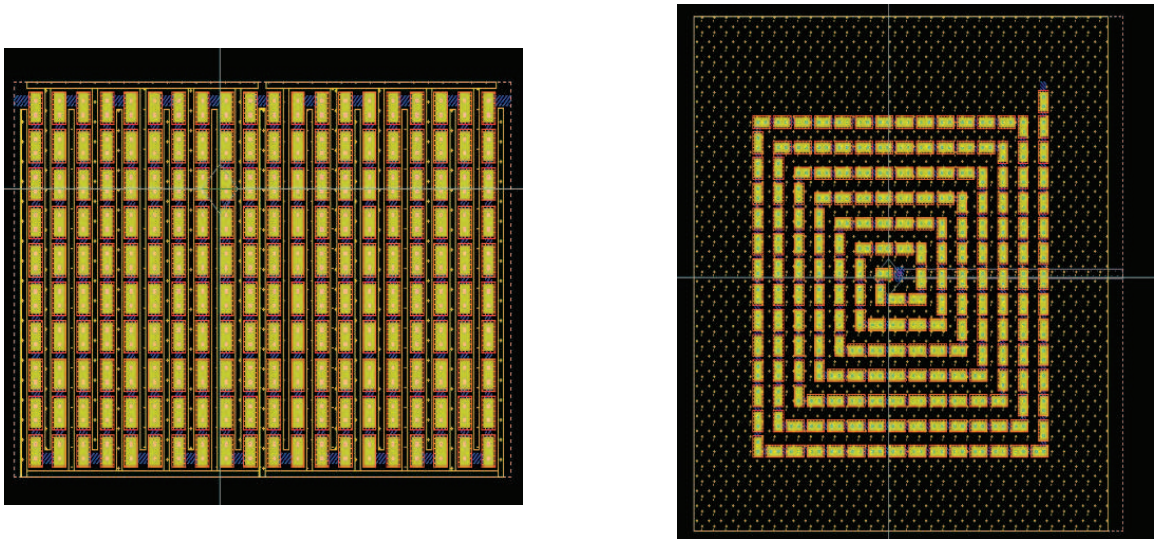


Figure 3.23: The two layout for a 400-JJ array. One is the serpentine structure with smaller inductance due to the flux canceling out, and the other one is the spiral structure with a larger inductance. For the spiral structure, if the inner end is connected to the output node, the delay will be decreased because of the smaller effective total capacitance.

node where the voltage change is highest, and the smallest contributor is the capacitance associated with the junction that is closest to the V_{DD} node. Considering that the capacitance is evenly distributed in the previous analysis, and taking this distributed effect into account, the discharge time should be about half of the delay time of previous analysis.

In the current design, a 400-junction array that is required in order to have at least 1 V output swing consumes an area 400 μm $\times$ 500 μm in a 2.5 kA/cm^2 process. Without removing the ground plane, this large array would introduce a large total parasitic capacitance of 10 pF. But to keep the amplifier delay below 100 ps, we need to reduce the load capacitance below 60 fF. We must apply some aggressive ground

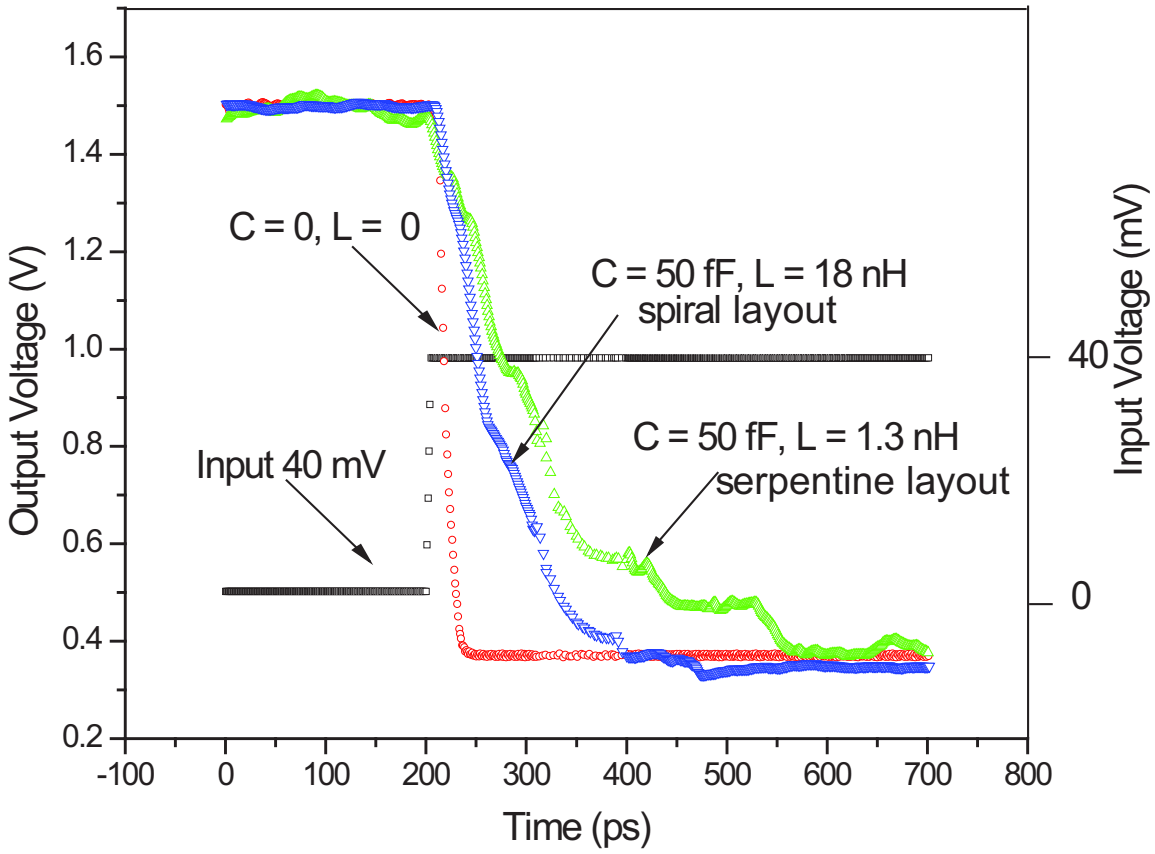


Figure 3.24: The simulation results confirm the qualitative analysis for the two junction array structures. The spiral one has smaller delay due to better capacitance distribution. But the potential problem is the antenna effect.

techniques to decrease the capacitance substantially, as discussed in previous section. Fig. 3.23 shows two layouts of the long junction array.

Fig. 3.24 a) is a serpentine array and b) is a spiral array. The serpentine structure has a smaller parasitic inductance because the magnetic flux in neighboring arms cancels out. The spiral array has a much larger inductance due to the flux enhancement. In terms of capacitance, they share the same total capacitance if the array is treated as a connected conductor. However, individually, the capacitance for the

inner junction is smaller than the capacitance for the outside junction. If we connect V_{DD} to the outside junction and the output node to the inner end, the discharge time will be further reduced. As we discussed before, part of the price we pay in order to have a spiral array is the increase of inductance. However, the large resistance of the cascode current source ensures that the increased inductance does not add too much in delay. The only practical problem for this spiral array is its greater potential for absorbing and transmitting high-frequency signals, such as the frequency components in the switching process. Fig. 3.24 shows the simulation results for these two structures. The capacitance value is calculated based on a 50 μm gap between junction array edges and the ground plane edges, and the inductance value is calculated based on Eq. 3.9. The simulation results confirm that the distributed model has a smaller delay and the spiral array can further improve the delay.

In designing the hybrid amplifier, decreasing the parasitic load capacitance is the key point to reduce delay time. With the proposed ways of reducing capacitance, less than 30 fF is possible, which would give a delay of 60 ps. Parasitic inductance, as well as the nonlinear junction inductance, contributes to delay time as well, but in a different and much weaker way.

3.5 CMOS memory core and peripheral circuits

The interface circuit amplifies an SFQ pulse to a volt-level digital signal that can be used by low-temperature CMOS digital circuits. As was discussed in Chapter 2,

CMOS digital circuits operate faster and consume less power at 4 K. The design of the memory core of a hybrid memory benefits from the low-temperature operation as well. In the following paragraphs, a CMOS-based DRAM memory system will be discussed and the focus will be the differences between operation at different temperatures.

In order to choose the right cell structure for a DRAM memory, it is useful to first compare different cells, both at room temperature and at liquid helium temperature. Fig. 3.25 shows the schematics of different memory cells that are commonly used in the semiconductor industry. Fig. 3.25 a) shows a standard 6-transistor (6-T) static random access memory (SRAM) cell. In this cell, information is stored by two cross-coupled inverters. This inverter couple has two stable states and an unstable middle point; due to noise, the unstable state cannot stay long. Therefore, the inverter couple stores either a zero voltage or a V_{DD} voltage at node Y. The read/write processes are carried out by the two NMOS transistors by carefully sizing them. This cell is called a static memory due to the fact that the information stored will not change as long as the power supply is on; the cross-coupled inverters keep the information regardless of how much leakage current there might be.

Starting from SRAM, if we take the two PMOS transistors out of the cell, as shown in Fig. 3.25 b), it can still work as a memory cell. However, since there are no such pull-up devices as the two PMOS transistors of Fig. 3.25 a), there is not a feedback mechanism to maintain the information stored in node X and node Y. Due to the subthreshold leakage path through M_1 and M_2, the charges on node X or Y

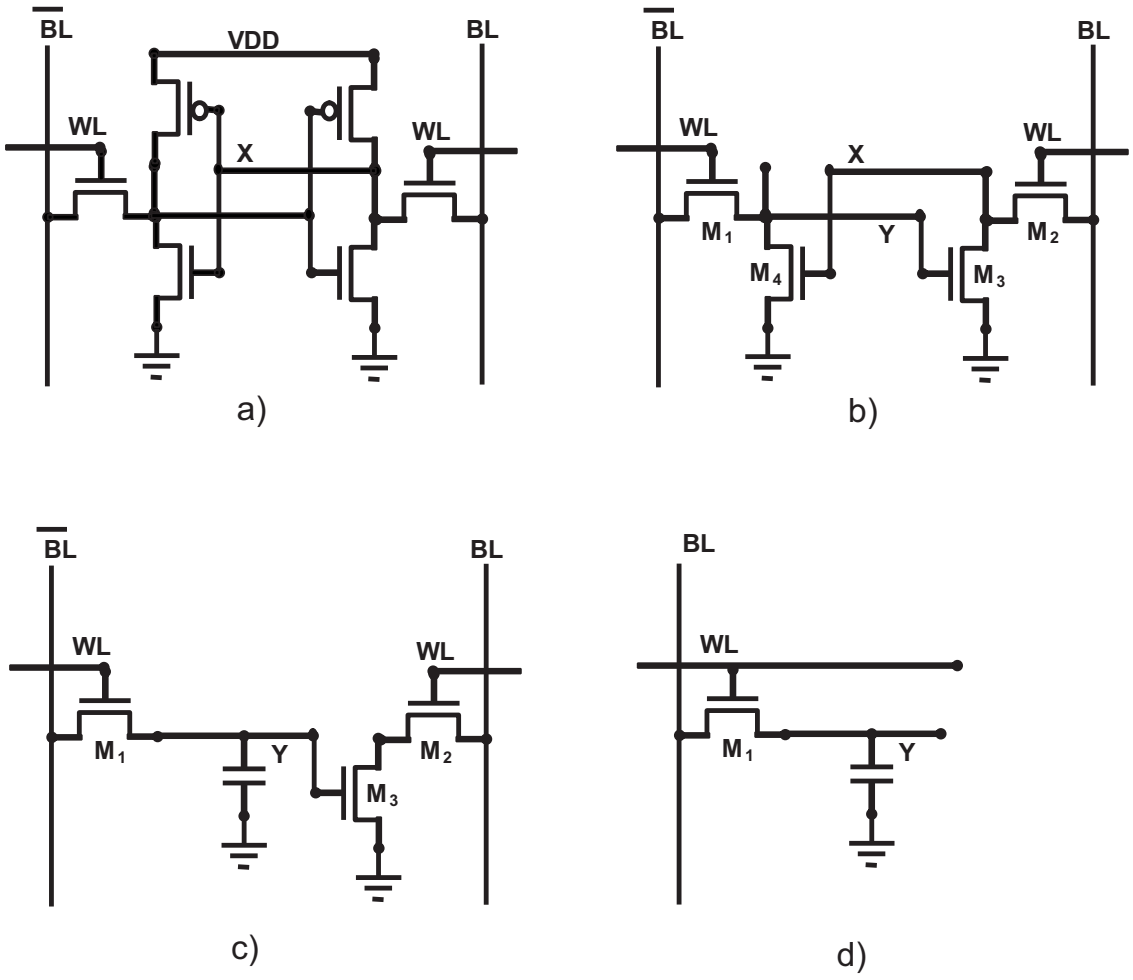


Figure 3.25: Standard memory cells in the semiconductor industry. a) 6-T SRAM cell. b) 4-T DRAM cell with differential operation. c) 3-T DRAM cell with a non-destructive read-out. d) 1-T DRAM cell (the capacitor can be implemented by deep trench to save area.)

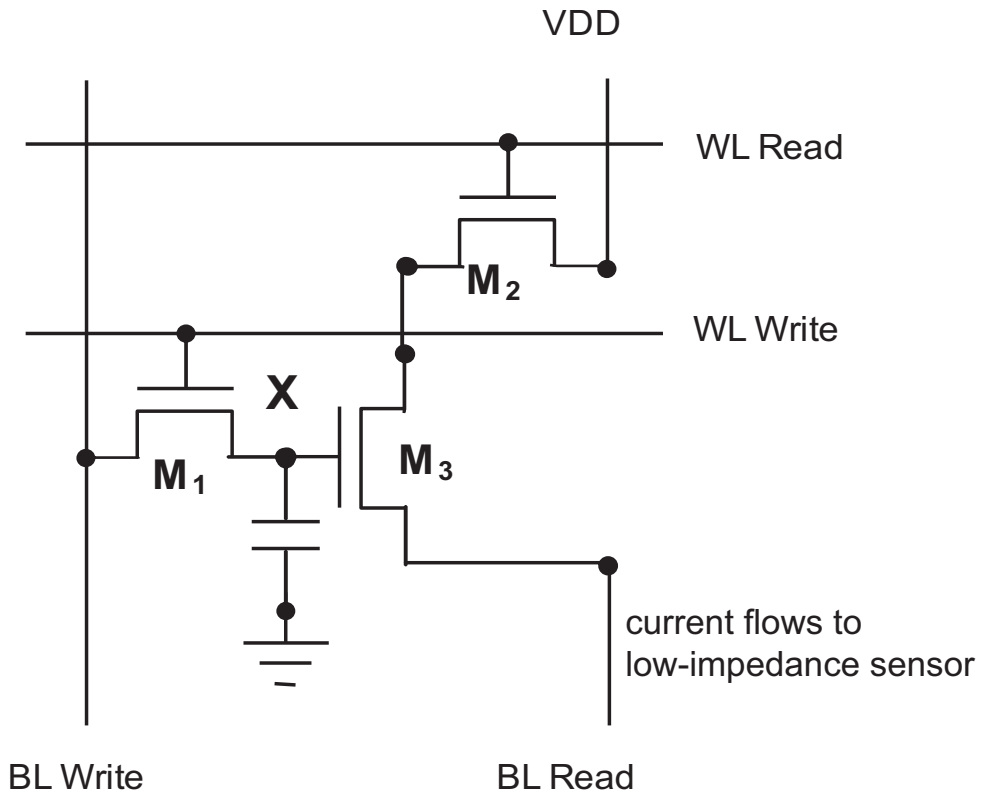


Figure 3.26: The 3-T cell in the hybrid memory system is different from the traditional 3-T cell. By connecting the bit line to the low-impedance current sensor, there is very little discharge delay time. So the total delay of a reading process can be reduced significantly.

will be discharged after some time, which typically is in the order of milliseconds. Therefore, some refresh circuitry is needed in order to retain the stored information. And this is why this 4-T cell is called a dynamic random access memory (DRAM) cell.

If we go further, we find that M_3 can be removed if differential operations (BL and $\overline{BL}$) are not required. Therefore, we turn a 4-T DRAM cell into a 3-T DRAM cell,

as shown in Fig.3.25 c). The parasitic capacitance at node X is enough to support charge storage. Note that the readout process is a non-destructive readout since the information is stored on a gate capacitor and is not drained upon reading.

If more density is required, we can further remove M_2 or both M_2 and M_3, leaving a 2-T DRAM cell or a 1-T DRAM cell. For a 1-T DRAM cell, a large built-in capacitor is required due to the charge-sharing problem. And the readout process is destructive, so the readout process is always followed by a re-writing process. In this dissertation, the 3-T structure is chosen because a nondestructive readout is preferred as well as is high density. However, the traditional 3-T cell is a voltage-sensing cell, which means that after the readout, the voltage on the bit line is sensed to determine whether a "0" or a "1" is stored. In our hybrid memory, current flows from the cell and is fed to a Josephson current sensor at the end of the bit line, which translates a current into an SFQ pulse that goes to the processor. Fig. 3.26 shows the present version of a 3-T cell. In this cell, the bit line is connected to V_{DD} and the current is fed directly into such a low-impedance current sensor that the source of M_3 is effectively grounded. During a writing process, M_1 is turned on and the capacitor is charged to $V_{DD} - V_T$, due to "threshold loss". In a reading process, M_2 is turned on and current flows through M_2 and M_3, to the current sensor, causing an SFQ pulse output. The size of M_2 and M_3 has to be designed carefully so that the reading current is within the input margins of the current sensor.

The low-temperature operation increases driving capability of transistors, and

reduces the parasitic capacitance, both of which lead to a faster writing and reading than those of the room-temperature operation. Also, due to the sharp subthreshold swing at 4 K, the subthreshold leakage current of M_1 is extremely small; therefore, the voltage at node X retains its value after a long time. The retention time, defined as the delay until the voltage on node X is less than half of the original value is increased to almost infinity. There is no need for refresh circuitry. In other words, the low-temperature operation turns a 3-T DRAM cell into an SRAM cell, or even better than an SRAM cell; this 4 K 3-T retains its datum whether there is a supply voltage or not.

Another important benefit brought by low-temperature operation is the reading time improvement. For a traditional room-temperature DRAM, in order to read a cell, the bit line has to be pre-charged by a large PMOS device. When the WL Read is turned on, M_2 is turned on and it starts to pull current from the pre-charged bit line. After some delay time, the bit-line voltage is sensed to tell whether a "1" or a "0" was stored. This delay time depends on how much charge is on the bit line capacitance and how large is the discharging current. The bit line has a large capacitance due to the fact that it is connected with many transistors. The delay is fairly large and increases with increasing memory size, due to increasing capacitance. The 4 K hybrid DRAM, on the other hand, has a very short read delay time. The bit lines are always connected to the power supply and the current pulled out from the power supply is directly fed into the low-impedance current sensor. The delay is extremely small,

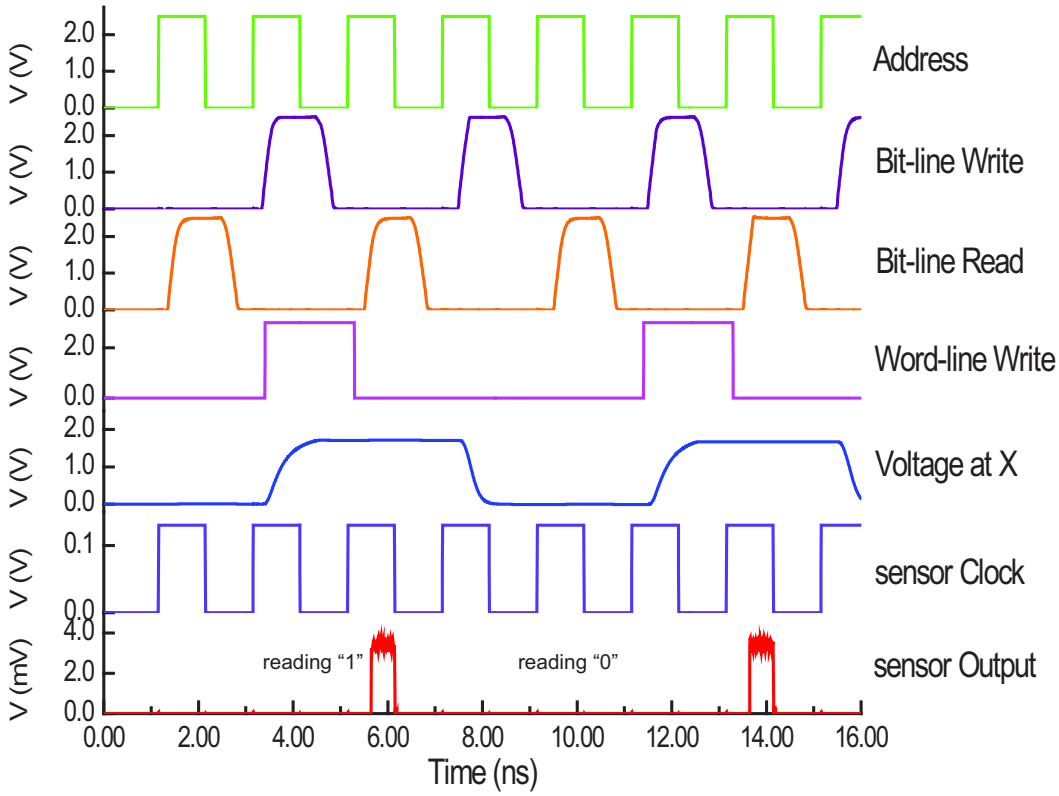


Figure 3.27: The simulation results of a reading process including address buffer and decoder, based on the 4 K CMOS model. The access time is about 400 ps, which is reduced by a factor of two compared to room-temperature operation. The improvement is contributed by both the low-temperature CMOS operation and the different reading scheme.

compared to the delay at room temperature.

The memory cells store charges to represent digital information. But to access them, an address decoder with buffers has to be carefully designed in order to achieve both minimum delay and minimum power consumption.

A decoder for a 64-kb memory is an 8-input AND gate with certain numbers of fan-out. In order to have less power consumption as well as more robustness, the "static" CMOS [41] logic family is chosen. In order to access the targeted cells as quickly as possible, the decoder is designed such that a low-to-high transition is faster than a high-to-low transition. And the design is based on the well-known "logical effort" method to obtain a minimum delay time [41]. Although the 4 K model is different from the room-temperature model, the low-temperature operation does not compromise the validity of this method. Fig. 3.27 shows the simulation of such a design, based on 4 K model established in Chapter 2. The simulation shows a 400 ps access time.

3.6 Performance conclusion

According to the above simulations, a total access time of 500 ps can be achieved in a 2.5 kA/cm^2 Nb process and a 0.25 μm CMOS process. The delay of the CMOS part is 400 ps and the delay of interface is 100 ps.

The total power consumption of the memory system differs for writing and reading because that determines how many interface circuits are used. For a typical writing

Table 3.1: Performance metrics for a 64-kb hybrid memory

Metrics	Values
Technologies	0.25 μm CMOS and 2.5 kA/cm^2 Nb
Read access time	500 ps
Operating frequency	1 GHz
Power consumption for read	10 mW
Power consumption for write	28 mW
Chip sizes	5 $mm \times 5$ mm and 2.4 $mm \times 2.4$ mm

process, 13 interface circuits are required for decoding and 32 interface circuits are required for data (if a byte of 32 bits is used, as proposed in the RSFQ 32-bit CPU project). And a typical reading process only requires 16 interface circuits. The dynamic power of the memory consumes only a small fraction of the whole power. For a 1 GHz operation, the dynamic power should be less than 100 μW. So the total power for a reading process is about 10 mW and the power for writing is 28 mW. Note that the power mainly is attributable to the interface static power. Many techniques can be utilized to reduce the static power, as will be discussed in Chapter 5.

Table 3.1 summarizes the performance details of such a 64-kb hybrid memory based on simulation, using a 2.5 kA/cm^2 Nb process and a 0.25 μm CMOS process.

Based on the simulation and optimization results, this hybrid memory is a strong candidate for a high-end RSFQ computing application.

Chapter 4

Measurements of 64-kb hybrid memories

In this chapter, we report on a 64-kb Josephson-CMOS hybrid memory fabricated using a 2.5 kA/cm^2 Nb process and a 0.18 μm standard CMOS process and the 4 K measurement results. The functionality is verified at low speed and the delay measurements of the hybrid interface circuits and memory are done at high frequencies. The bit-error rate test is being designed and will be done in the future. All the measurements proved the feasibility of the hybrid memory idea as well as the sub-nanosecond access time.

4.1 Test set-up

For testing room-temperature semiconductor circuits, chips are mounted on a PCB board which communicates with a PC and makes automatic measurements possible. However, our hybrid memory chips must be cooled to extremely low temperature. The chips are mounted on a low-temperature probe that is easily put into liquid helium. For this dissertation, a home-made probe with magnetic field shields is used for the low-speed testing. The chip is wire bonded to the chip holder of the low-speed probe, and the signals are connected to the measurement cables via regular wires. Therefore, the bandwidth of the probe is very low, only on the order of 100 KHz. However, when the chips are required to work at high frequencies, it is essential to use a suitable wide-band, low-temperature probe. In this work, all high-speed testing was done using a so-called Petersen probe made by America Cryoprobe, which is a commercial product and is shown in Fig. 4.1. The designer of this probe described it

Figure 4.1: The pictures of the 24-pin Petersen probe.

in detail in his Ph. D. dissertation [57] and a summary of it follows: The bandwidth of the probe is 12 GHz and on it one can mount one 5 mm $\times$ 5 mm chip that has 24 pads. In order to suppress the ambient magnetic field, there are two concentric shields; the shields can decrease the magnetic field by two orders. 24 high-frequency transmission line cables are used in order to transmit high-frequency signals between room temperature and low temperature and 24 microstrip transmission lines on a pc board. Spring fingers are used to connect to the pads on chips. The microstrip lines are designed such that the impedance is identical with that of the cables and the reflection are thereby minimized.

The Pertersen probe is widely used in superconductor laboratories for high-speed digital circuit measurements. In order to test our hybrid chips, a square hole in the center of the probe head had to be made and the reason follows. Because it is a Josephson-CMOS hybrid system, and there is no technology available to us to put the two technologies on a single chip, some bonding technique had to be applied to make physical and electrical connections between the semiconductor circuit and the superconductor circuit. In this work, the size of the superconducting chips is 5 mm $\times$ 5 mm, while the CMOS chips have a size of 2.4 mm $\times$ 2.4 mm. Two bonding techniques were used to put them together. The easiest bonding technique is wire bonding. For a wire bonding chip set, the CMOS chip is glued face up on the top center of the superconductor chip, and short aluminum or gold wires are used to connect pads on both chips, as shown in Fig. 4.2. This method has a obvious drawback of having

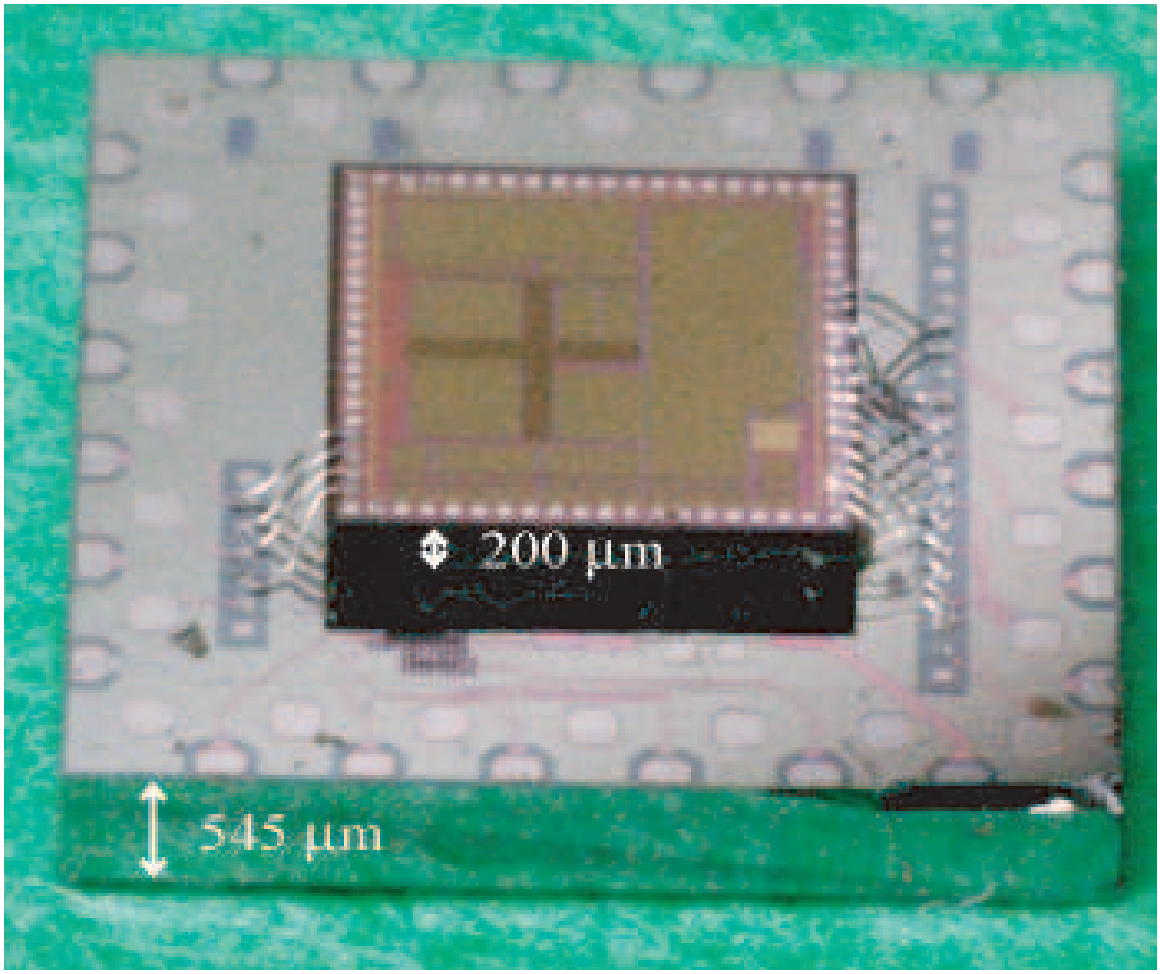


Figure 4.2: The photograph of a wire-bonded hybrid memory chip set. The CMOS chip was thinned to about 200 μm in order to reduce the length of the bonding wires and, therefore, the parasitic inductance.

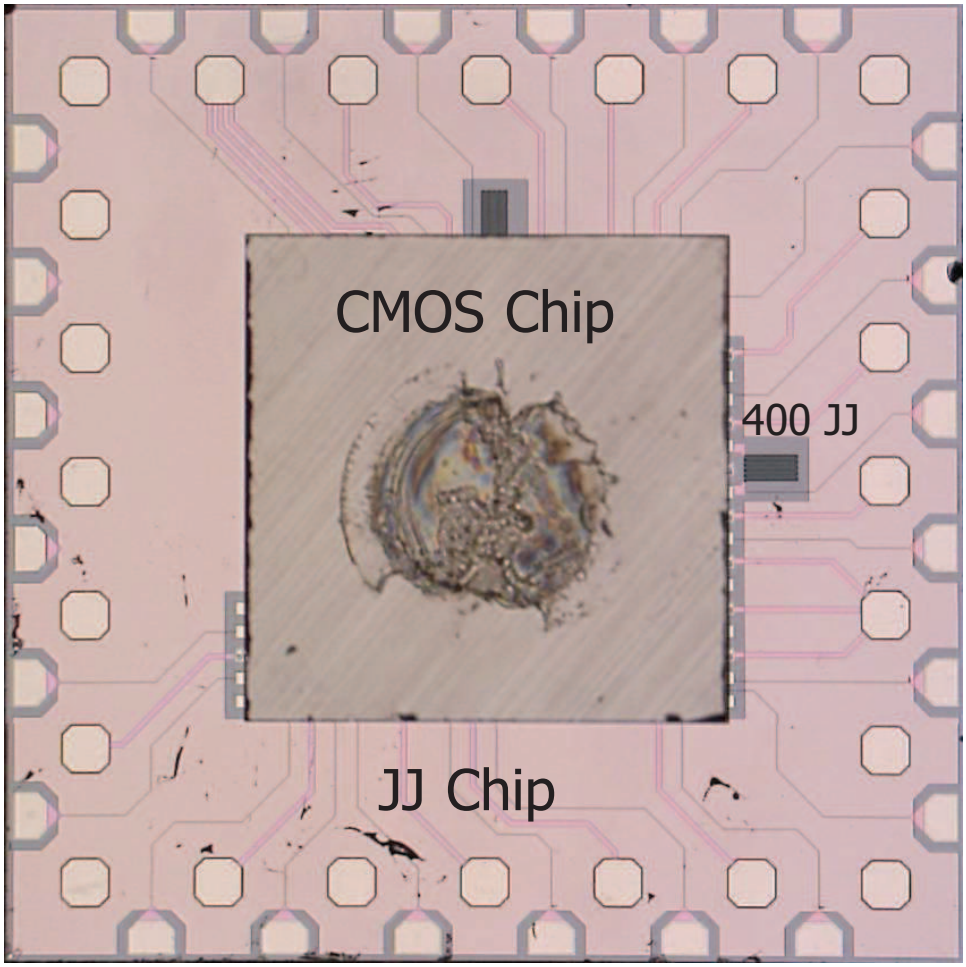


Figure 4.3: The photograph of a flip-chip bump-bonding memory chip set. There are two 400-junction arrays shown on the JJ chip.

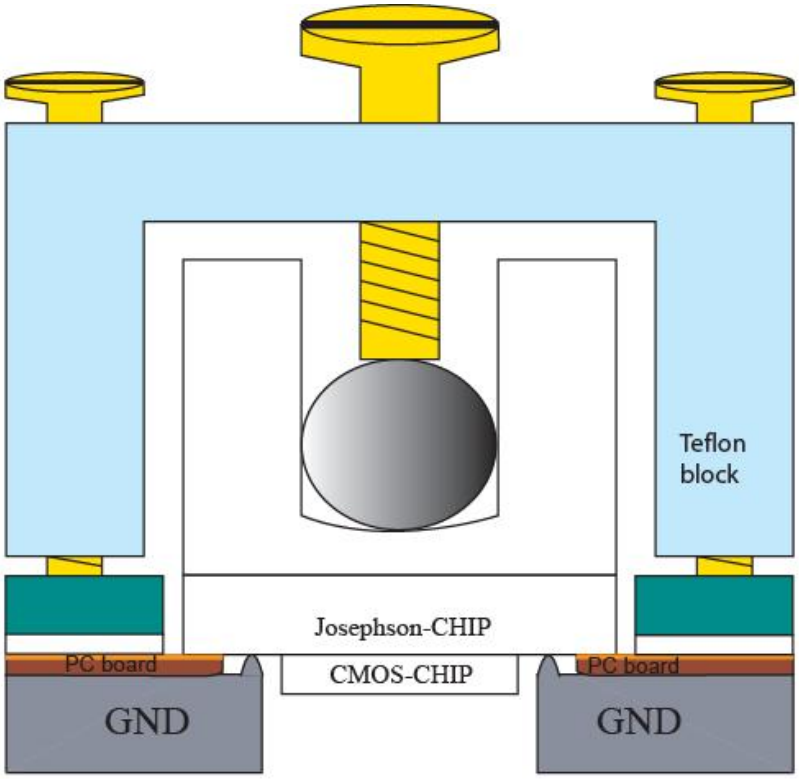


Figure 4.4: The modified Petersen probe with a square hole in the ground disk to accommodate the CMOS chip.

large wire inductance that apparently compromised our high-speed measurements. The other alternative was solder bump bonding. For a bump-bonding chip set, the both chips have a layer of gold on top of the pads. A solder alloy adheres to the gold when the chip is dipped in it. Then the CMOS chip is flipped face town onto the superconductor chip; the two chips are aligned with the help of a bonding machine, shown in Fig. 4.3. Both methods involve a chip set that has a smaller chip sitting on top of a 5 mm $\times$ 5 mm chip. Therefore, a square hole was cut in the disc so that the smaller CMOS chip can be accommodated when the pads on the superconductor chip make contact to the spring fingers, as shown in Fig. 4.4.

4.2 Flux trapping and magnetic shielding

Magnetic flux trapping is a critical issue for all Josephson circuits. On superconducting chips, the Nb is deposited on the silicon forming a thin film. If the magnetic field is parallel to the surface, it is easy for the flux line to find ways to get around the Nb film. However, if the magnetic field is perpendicular to the surface, it is hard in terms of energy for the flux line to find ways to get around the Nb film. Therefore, it is possible that some flux is trapped, causing significant negative effects in the superconducting circuits.

A single flux quantum passing through the junction could cause its "effective" junction critical current to change dramatically. Also a flux quantum within the area of a circuit loop would have the same effect as a logic quantum stored and, therefore,

compromise the performance of the circuit. The trapped flux are quantized as follows,

$$\Phi = n\Phi_0,\ n\ is\ an\ integer \tag{4.1}$$

Assuming the ambient magnetic field is only the earth's field (about 0.5 gauss) after a good quality shielding which can suppress the field by two orders of magnitude to 0.005 gauss, a small area of 100 μm $\times$ 100 μm can hold 240 flux quanta!

There is a circuit design way to minimize flux trapping. Certain areas of ground plane can be removed when designing the circuit, making the circuits protected by these "moats" surrounding the circuit. [56] In our design, the ground plane underneath junctions are removed in order to minimize the parasitic capacitance and this removal helps to prevent flux trapping.

In our experiments, serious shielding procedures have been taken and a process called "de-fluxing" has to be carried out if there is still flux trapped after shielding protocols are followed. The de-fluxing process is to take the probe out of liquid helium and make sure the temperature of the chips is higher than the critical temperature of niobium and then slowly put the probe back into liquid helium again, without any current or voltage source connected to the probe. Typically, de-fluxing several times can decrease the trapped flux to an acceptable level.

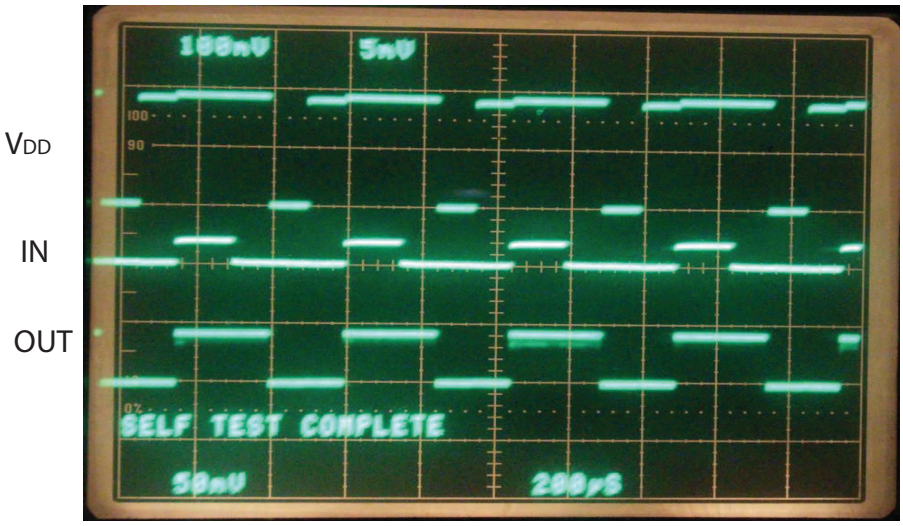
Regardless of how carefully the circuit is designed and the testing is done, flux trapping can happen. This problem is one of the most troublesome practical problems in superconducting circuit measurements. In the next a few sections, we will show the measurements with and without flux trapped.

Another practical issue with high-frequency testing is the electrical connection between the chip to be measured and the Petersen probe. There are supposed to be 24 ground pads made of copper bumps on the probe. However, after many cycles of remounting, not all the ground connections are perfect, some of them may be totally disconnected. It is not a problem for low-frequency measurements. For high-frequency measurements, however, disconnected ground pads can increase the loop of returning currents, so the parasitic inductance is increased. The high-frequency performance may be degraded by the larger parasitic inductance. A further probe problem is that the contact resistance of signal lines can lead to crosstalk between two pads.

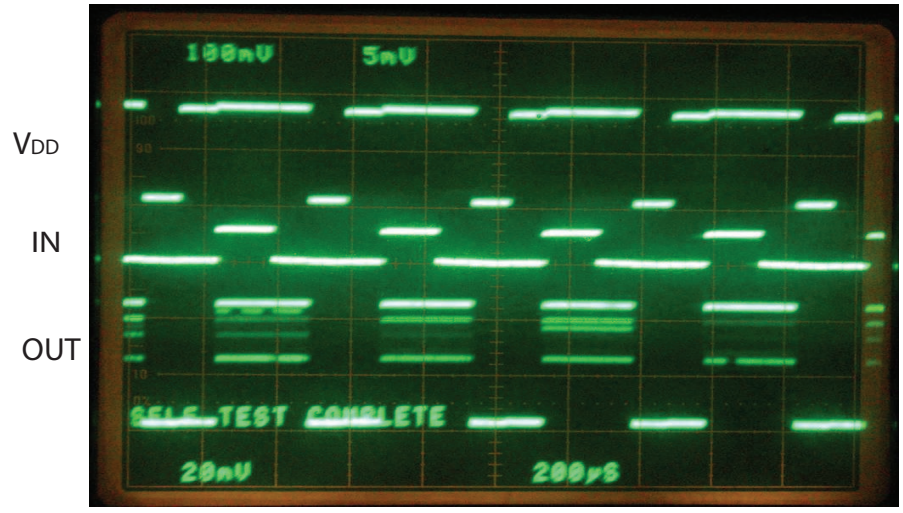
4.3 Low-frequency functionality test

4.3.1 Functionality test of the interface circuit

Fig. 4.5 and Fig. 4.6 show the low-frequency experimental results for the Suzuki stack and the second part of the hybrid amplifier, respectively. Fig. 4.5 (b) shows the waveforms when there was flux trapped. The clear waveforms in Fig. 4.5 (a) are the results after several times of de-fluxing. The presence of trapped flux makes the junction seem to have a smaller critical current, so the original bias condition does not apply. Some of the junctions may be biased close to their critical currents, causing the unstable waveform shown in Fig. 4.5(b).



(a)



(b)

Figure 4.5: The low-speed functionality test of the 2×16 JJ Suzuki stack. (a) No flux trapped (b) Flux trapped. The output is switched by the input signal and reset by the clocked V_{DD} signal. We attribute the multilevel output for flux trapping. The scales are, 100 mV/div, 5 mV/div, and 20 mV/div, for V_{DD}, input, and output, respectively.

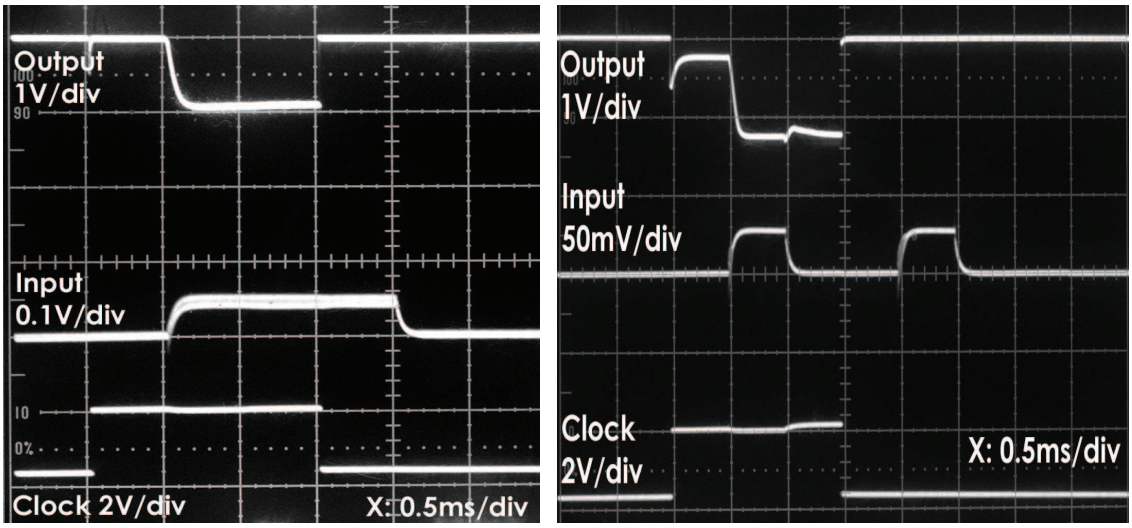


Figure 4.6: The low-speed functionality test of the second-stage amplifier. There is clock feedthrough for the clock with a smaller rise time (b) and no clock feedthrough for the clock with a larger rise time (a).

The Suzuki stack and the second-stage amplifier have been proved to work. Using a large coupling capacitance (10 pF), the 40 mV Suzuki stack output is transferred to the next stage. The parasitic capacitance associated with the coupling capacitor is very critical. (The parasitic capacitance is mainly between the output side of the coupling capacitor and the ground plane; efforts have been made to minimize it.) If the parasitic capacitance is too large, the capacitive voltage divider can degrade the voltage transferred to the gate of the bottom transistor. In previous designs, when the two circuits are combined together to test, the second-stage amplifier output swing is only 0.3 V. The reason was soon found out: the large electrostatic-discharge (ESD) protection pad capacitance and coupling capacitor make a voltage divider which lowers the voltage to the next stage: the coupling capacitor is 10 pF, made

by multiple metal layers forming parallel plate capacitors. However, one end of the capacitor is connected to the NMOS gate via a wire inductor and a pad capacitor. The I/O pads on CMOS chips have ESD protection circuits in order to protect the gate from being damaged by static charges. Since the voltage caused by static charges may be as large as several thousand volts, the protection circuits have to provide a very large current-delivering capability to remove the charges. Therefore large MOSFETs are used in the protection circuits, leading a large capacitance to ground. In our later designs, the protection circuits are removed and we get the full voltage swing.

4.3.2 Memory cell functionality tests

Fig. 4.7 shows the functionality test of a memory cell at 4 K. The reading and writing operations are shown in the picture. More important is that the retention time of the memory cell is almost infinity at 4 K, due to the sharp subthreshold swing, as discussed in Chapter 2. Our collaborators at Yokohama National University, measured the memory retention time at both room temperature (300 K) and liquid helium temperature (4 K) [42]. The 3-T memory cell's retention time at 300 K is about several seconds while the retention time at 4 K is more than 24 hours according to their experiments (no degradation was observed). In order to get a basic idea how long the retention time will be at 4 K, retention times at different temperatures are measured and plotted on Fig. 4.8. By extrapolating back to 4 K, an estimated retention time of 10^{482} years is expected for this 4 K 3-T DRAM cell.

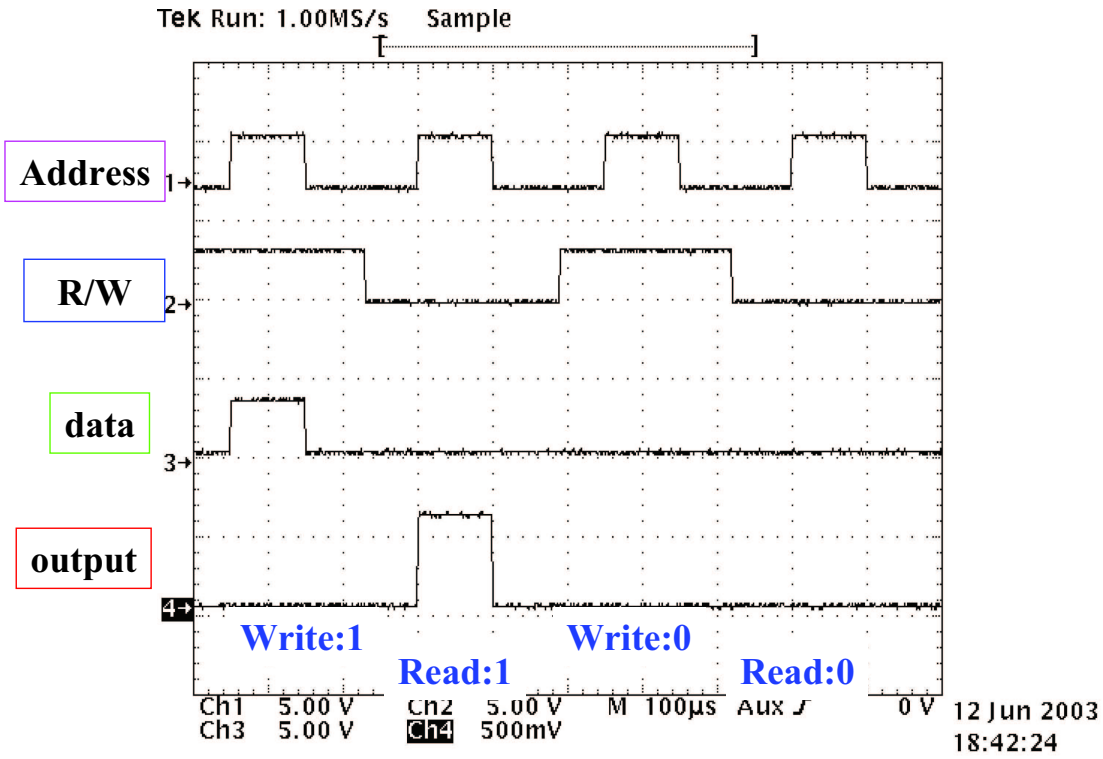


Figure 4.7: The low-speed functionality test of the memory core at 4 K. The signals are all CMOS volt-level signals.

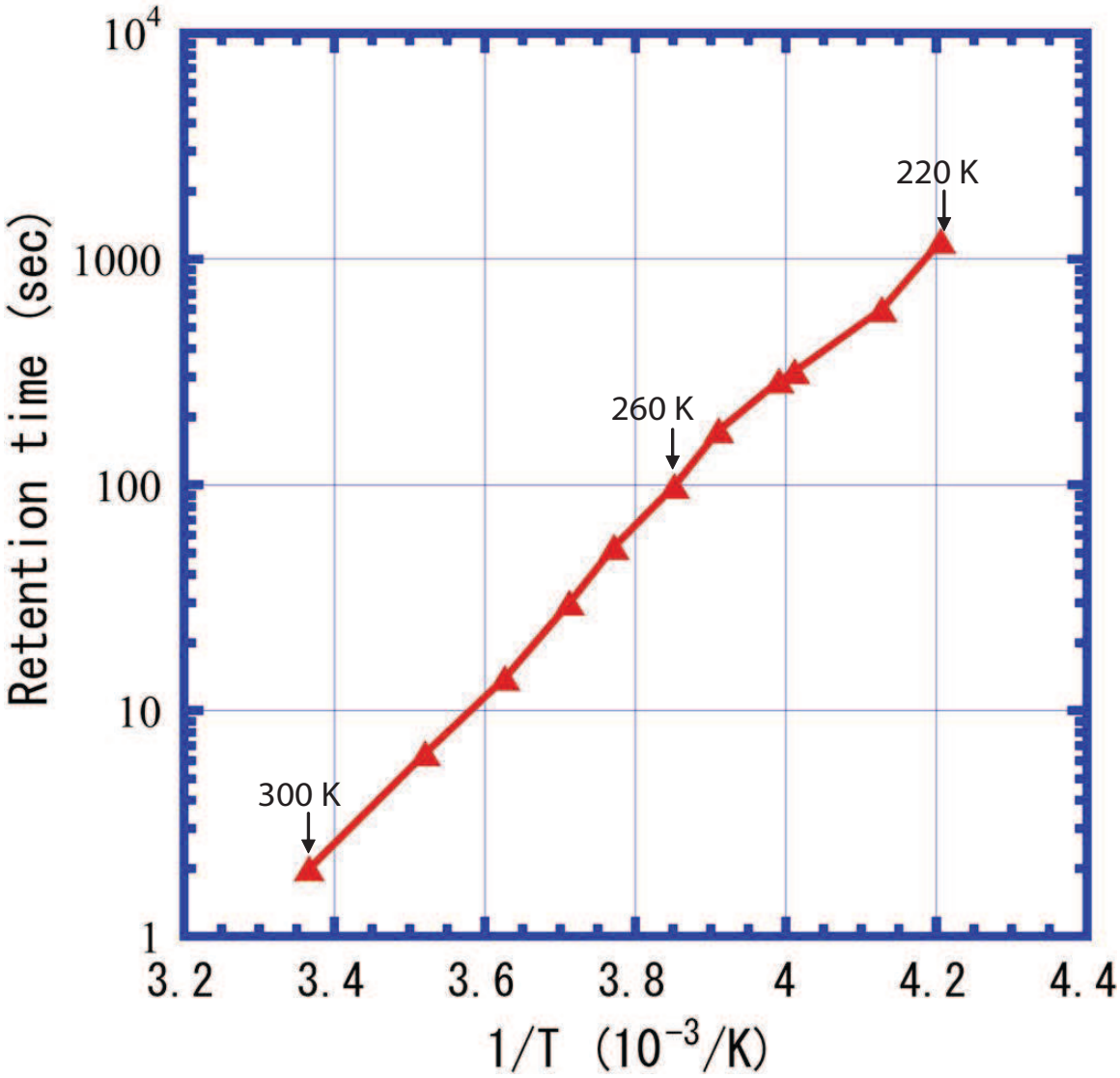


Figure 4.8: The retention-time measurement results at different temperatures. The 4 K retention time is believed to be 10^{482} years according to the extrapolation.

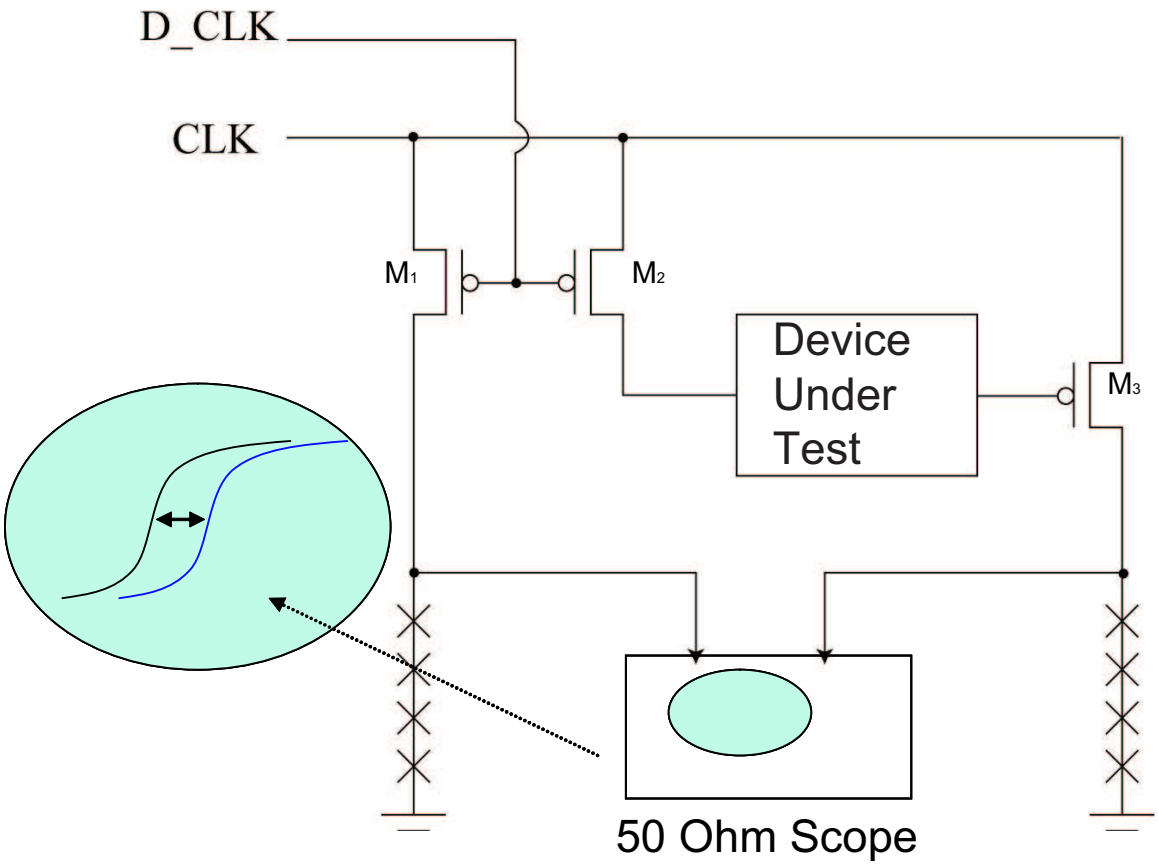


Figure 4.9: The delay measurement circuit for small delay measurement. The circuit under test can be interface circuit, the memory core, or the whole critical path. The precision of the measurement depends on the cable length precision and is measured to be less than 20 ps.

4.4 High-frequency test of the system

4.4.1 Measurement circuit and testing set up

Simulations have shown that gigahertz operation of the hybrid memory is possible and sub-nanosecond access time is expected. However, it is difficult to measure such small delay time in traditional ways. In the semiconductor field, it is common to use ring-oscillators and to measure the frequency of the oscillator; the delay time

of an individual inverter is then obtained from the oscillation frequency. This idea, however, does not apply to our memory and interface measurements, due to the large required area. U. Ghoshal [53] proposed the hybrid circuit in Fig. 4.9 to measure directly small individual delays.

The MOS devices M_1 and M_2 are designed so that the "ON" current is high enough to switch the 4-Josephson-junction (4JJ) arrays as well as the Suzuki stack in the device under test (DUT). The clock (CLK) and the delayed clock (D_CLK) are provided by an external clock generator with a variable delay delay device. The MOSFET M_3 is identical to M_1 and M_2 so that any parasitic effects will be compensated in the measurement. At some time t_0, both M_1 and M_2 are triggered by the clock signals (when CLK is high and D_CLK is low), after some small time delay (t_p) caused by the parasitic capacitance, the left 4JJ is switched and OUT_1 is about 10 mV; right after the M_2 is turned on, it delivers enough current to drive the interface circuit. After the delay of the circuit, M_3 is turned on and after the small time delay of t_p, the OUT_2 is switched to about 10 mV. The time between OUT_1 and OUT_2 is, therefore, the delay of the interface circuit. For the delay measurement of the memory decoder and memory cell, the same circuit applies with some minor modifications. Since the output of a hybrid memory is a current reading out from the cell to the input of a Josephson current sensor, we can simply connect the output node of the memory cell to the 4JJ, as long as the parasitic capacitance is compensated such that $C_{OUT1} = C_{OUT2}$. So the time difference between OUT_1 and OUT_2 is the delay of

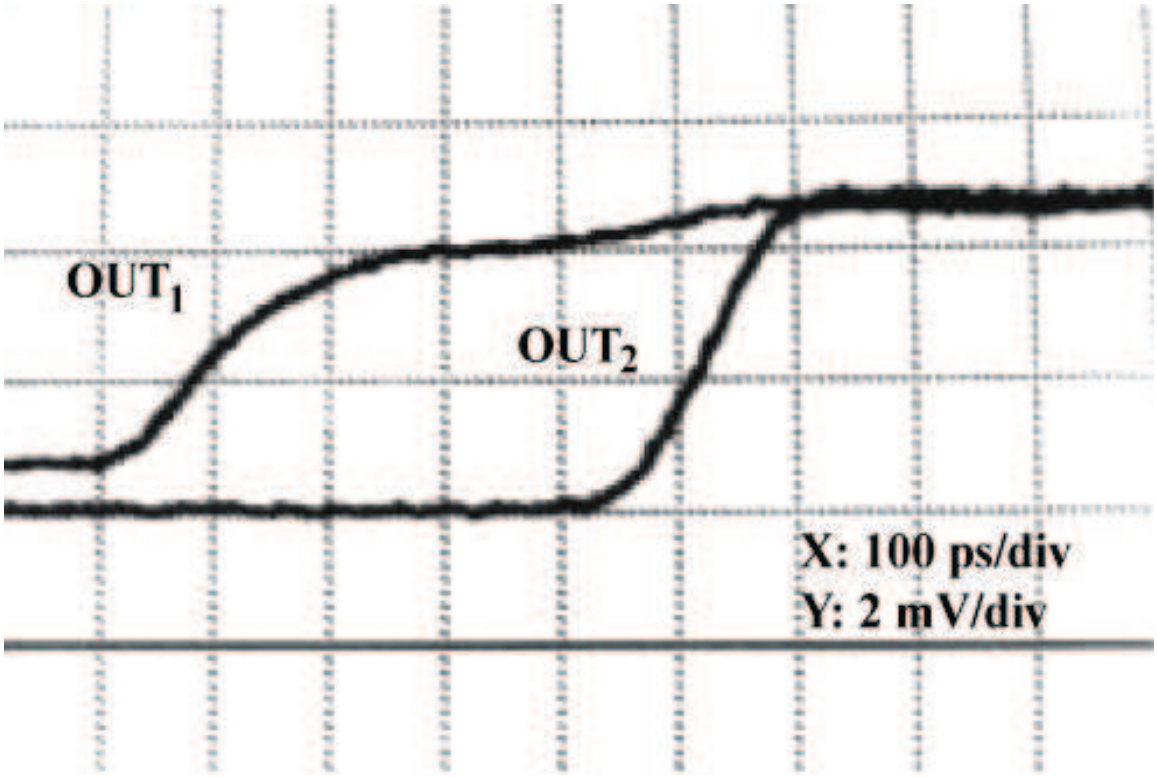


Figure 4.10: The delay of the second-stage amplifier. 430 ps measured delay is larger than the simulation results, which is explained in the text.

the memory.

The two cables from OUT_1 and OUT_2 to the oscilloscope must be exactly the same in type and in length. With careful set up, the accuracy of this measurement is believed to be about 20 ps.

4.4.2 Interface-circuit delay measurement

Fig. 4.10 shows the delay-measurement result for the second stage of the hybrid interface circuit, for a wire-bonded chip set. From the picture, one can read a 430-ps delay. This value is larger than the 170 ps found from simulation. The possible

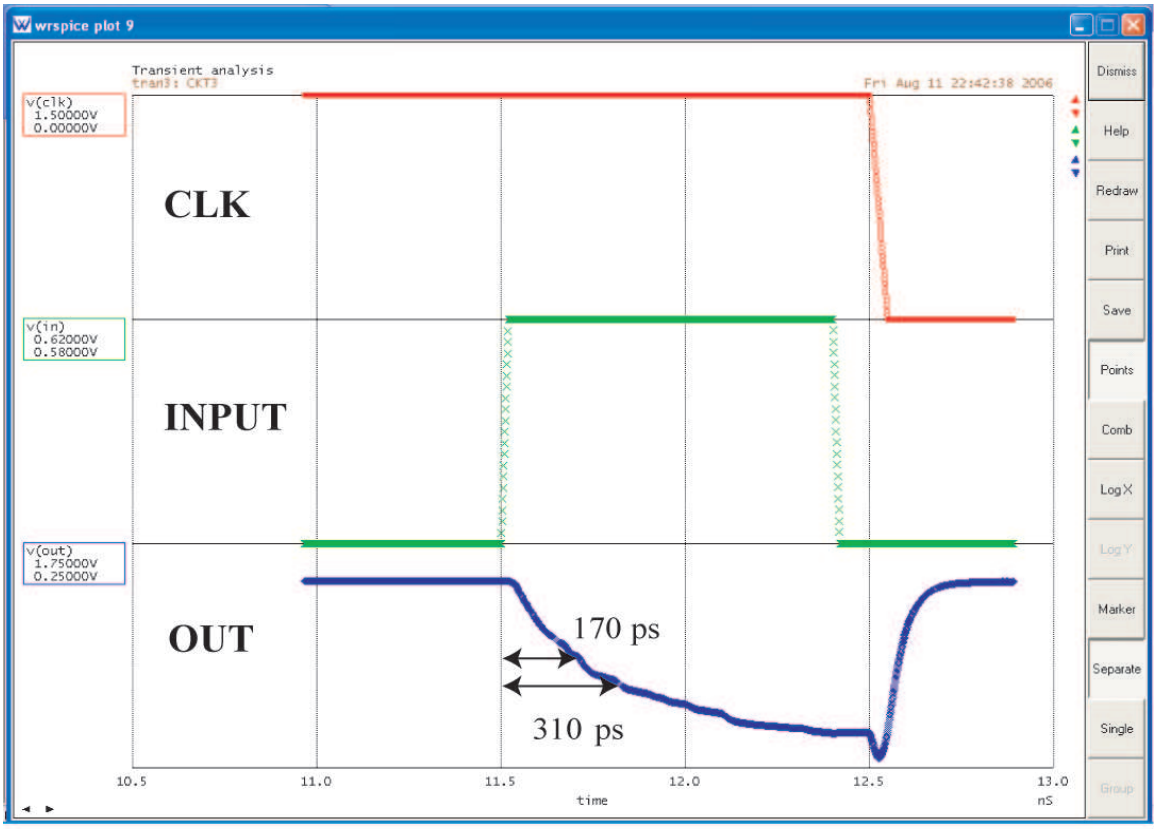


Figure 4.11: This simulation result for the second-stage of the interface amplifier shows that a large delay (310 ps) is incurred in obtaining the necessary 0.7 V to drive the next stage.

reasons follow: The delay time simulated previously was based on the difference between the half-level points of the input and output. But half of the voltage drop of the interface circuit, 0.5 V, is insufficient to drive the following PMOS with high enough current to switch the 4JJ output array. Rather, 0.7 V is required. Fig. 4.11 shows the simulation curves for the second stage of the interface amplifier where it can be seen that the delay would be 310 ps to obtain a 0.7 V drop. Thus 140 ps of the difference (260 ps) between the measurement and simulation is accounted for. Also, unidentified parasitics from the bonding wires or pad connections may increase the delay time.

Fig. 4.12 shows the delay measurement waveforms of the interface circuit (including the Suzuki stack and the second stage amplifier) for a bump-bonded chip set. The interface circuit is biased with a smaller V_{DD} (1.4 V) than the CMOS V_{DD} (1.5 V). Due to the leow parasitic bonding inductance and different bias condition, the measured delay 200 ps is smaller than for the wire-bonding chip set.

One can expect that a high-current-density JJ process will decrease the delay time of the interface amplifier due to the smaller junction capacitances. Simulations show that 20 kA/cm^2 Nb process will decrease the delay time to less than 100 ps.

Furthermore, we can take advantage of the sharp subthreshold property of 4 K MOSFETs and make the PMOS M_3 V_{DD} 0.3 V higher than the interface V_{DD}. When the interface circuit is off, this 0.3 V drop (V_{SG}) is not large enough to turn the following PMOS on, which would cause a problem, however, when the interface circuit

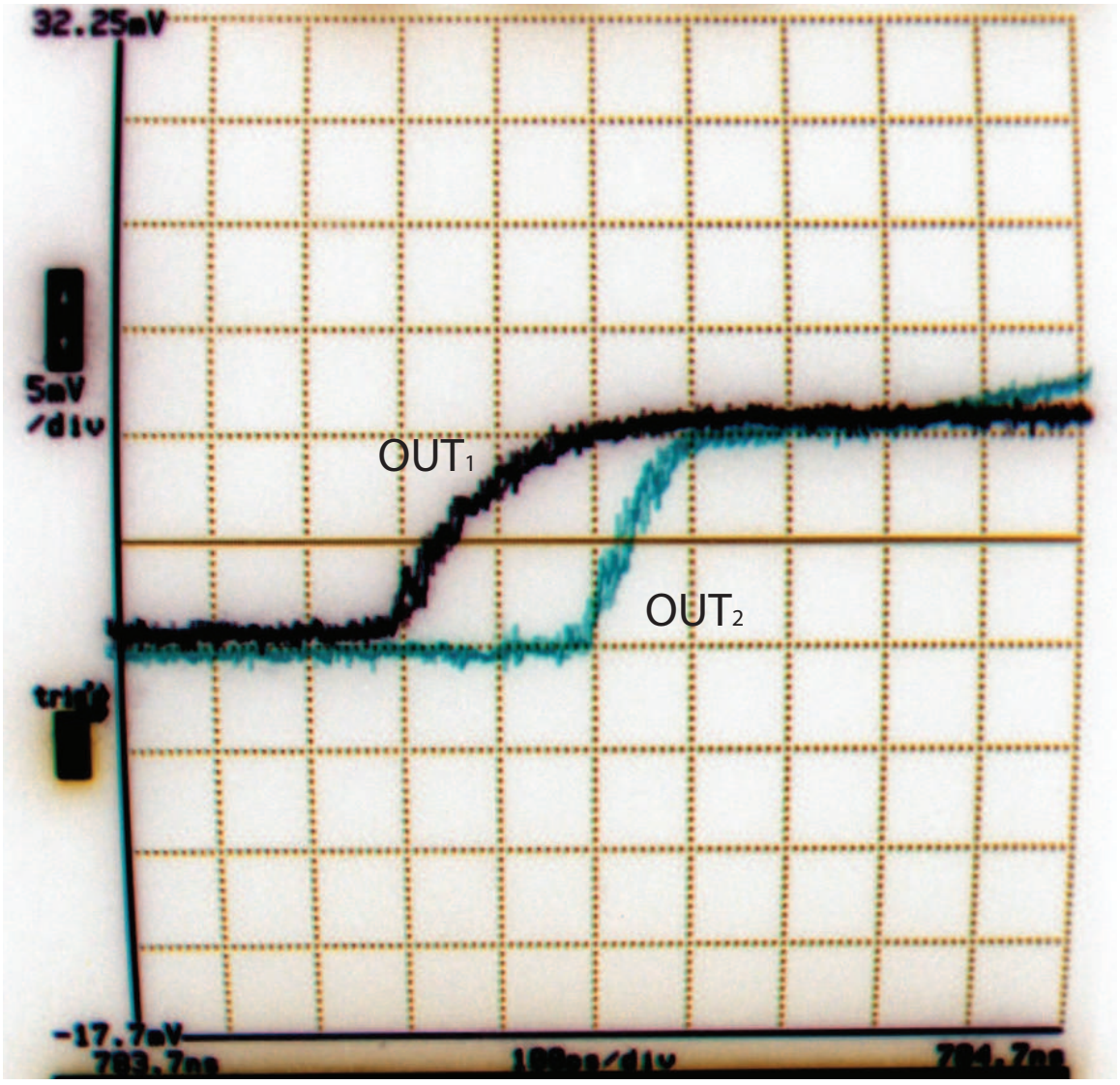


Figure 4.12: The delay of the second-stage amplifier, measured from a bump-bonded chip set. A 200 ps measured delay is smaller than the one that was measured from a wire-bonded chip set. X = 100 ps/div, Y = 5 mV/div.

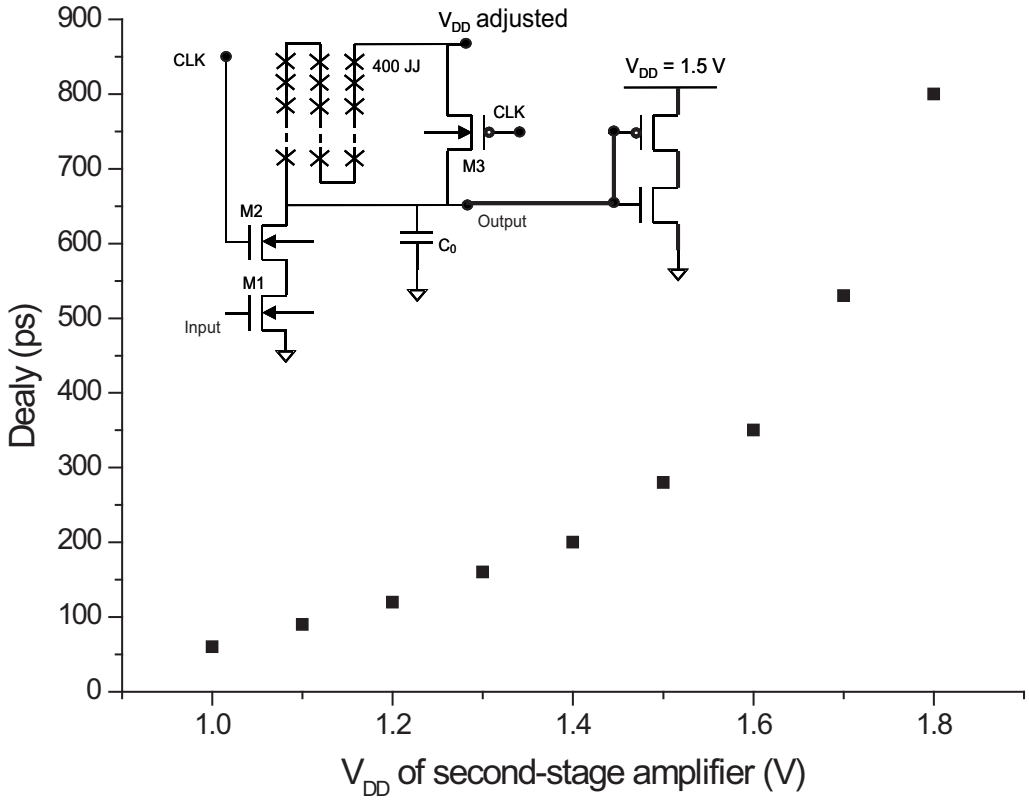


Figure 4.13: The measured delay time of a second-stage amplifier versus supply voltage of the next stage.

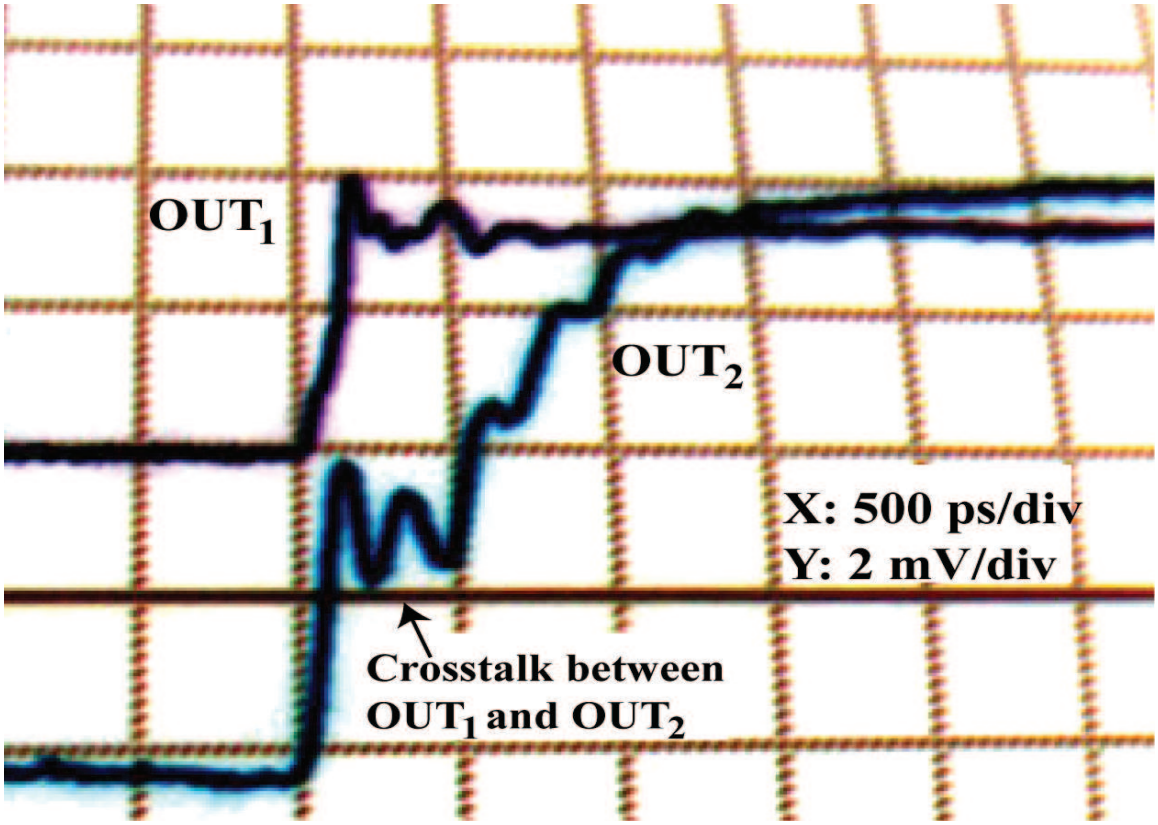


Figure 4.14: Memory delay measurement waveforms including input CMOS driver, decoder, memory cell, and bit-line JJ readout. About 500 ps delay time is measured, with $V_{DD} = V_{CLK} = 1.5\ V$.

switches, the 0.5 V drop in 400 JJ is equivalent to a 0.8 V drop for the following PMOS. Fig. 4.13 shows the measured relationship between the two $V'_{DD}s$ and the delay time.

By doing this, the delay time would be less than 100 ps for higher JJ current densities.

4.4.3 Memory-core delay measurement

In order to measure the delay of the CMOS part of the memory, the same testing circuit has been applied to the memory as was described in Sec. 4.4.1. The result is

shown in Fig. 4.14. A 500 ps delay is displayed in the picture. The supply voltage of the CMOS circuits is 1.5 V. We also observed that the delay decreased when we used higher supply voltage for the CMOS circuits.

The total access time is the delay of the interface circuit plus the delay of the CMOS memory part. Based on the test results obtained separately, we can conclude that the total access time is less than 600 ps. It could be better if we measured it directly. The measurement for the whole signal path has not been possible at this time, mainly due to the small margin of the second-stage amplifier and the clock crosstalk. But measurement results for the signal path from second stage of the interface amplifier to the memory readout current are obtained and shown in Fig. 4.15. The delay shown in the figure is about 600 ps. Taking the Suzuki stage delay time into account, the total access time is about 620 ps, which confirms the previous access time conclusion based on the interface delay and CMOS memory delay.

4.5 Discussion and conclusions

In this chapter, we demonstrated a 64-kb CMOS-Josephson hybrid memory operating at 1 GHz and at 4 K. Each individual part of this 64-kb hybrid memory system has been measured, both functionality and delay time has been verified and reported. Although a full critical path measurement is not possible at this time, multiple section measurements provide strong evidence that the total access time can be well under 1 ns, which fits the simulation results well. Power measurements also fit the calculated

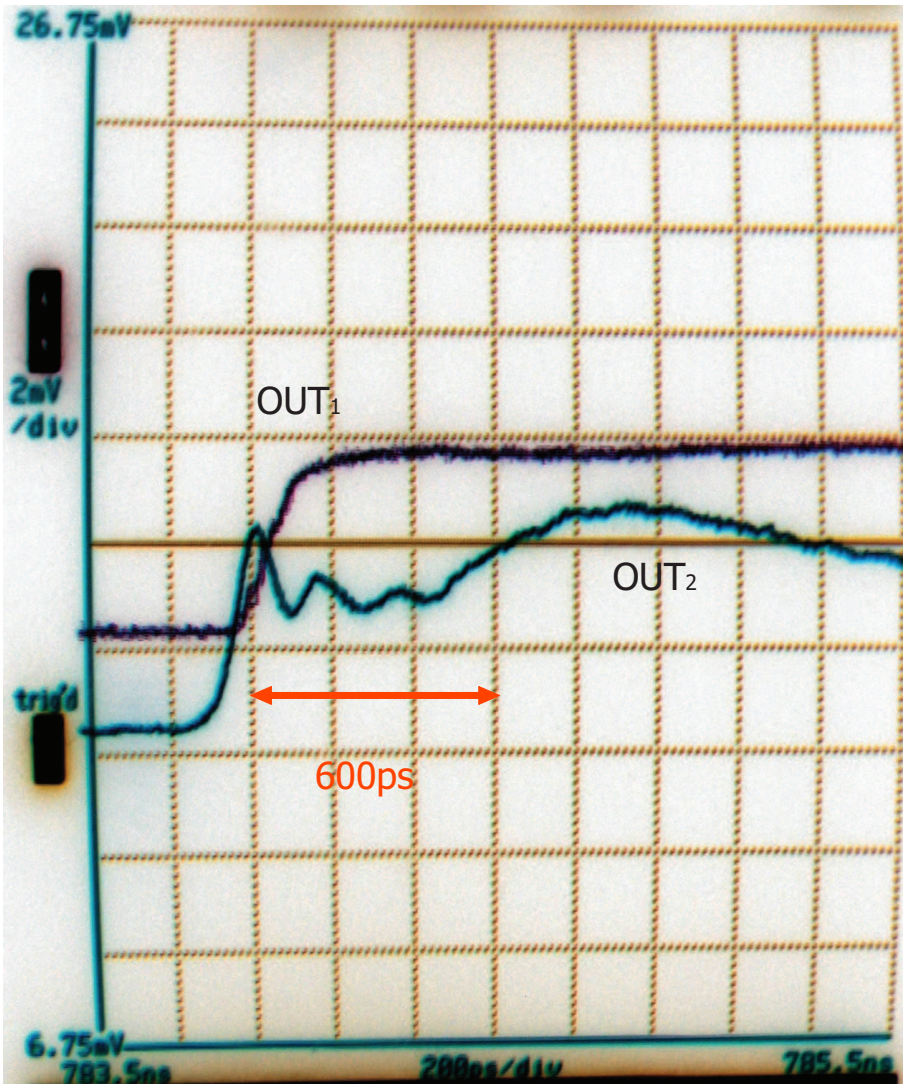


Figure 4.15: Delay measurement waveforms including the second-stage amplifier, input CMOS driver, decoder, memory cell, and bit-line JJ readout. A delay time of less than 600 ps is measured, with $V_{DD} = V_{CLK} = 1.5\ V$

value for the interface circuit as well as the memory core.

With more advanced CMOS and Josephson technologies in the near future, we strongly believe such a CMOS-Josephson hybrid memory can be expanded to larger sizes, with even faster access times, working at higher frequencies, and make a good candidate (probably the best one), for the solution of the long-standing memory bottleneck problem faced by superconducting digital electronics.

Chapter 5

Discussion and Conclusions

5.1 More advanced technologies for 64-kb hybrid memories

The previous chapters reported a 64-kb Josephson-CMOS hybrid memory that was designed and demonstrated using a 0.25 μm CMOS process and a 2.5 kA/cm^2 Nb process. With the continued development of both semiconductor and superconductor technologies, more advanced processes are expected to be used in the hybrid memory in the future. A 20 kA/cm^2 Nb process has been demonstrated in the laboratory [58] and 90 nm CMOS processes are now industrial standard. Based on the scaling rules of CMOS and assuming that the demonstrated low-temperature improvement of CMOS does not change much with scaling, we can estimate the delay and power consumption for a 64-kb hybrid memory using advanced technologies. Since the delay comes mostly from the CMOS part and most power dissipation is the interface static power (proportional to V_{DD}), one can conclude that the upgrade of the CMOS process will affect the performance of hybrid memory more than the upgrade of the Josephson process. Scaling-rule calculations suggest that the access time for 90 nm CMOS will be 240 ps while power consumption decreases because V_{DD} decreases. Since the interface static power dominates, the power reduction factor is about 40%, if a 0.9 V supply voltage is used.

We must point out that the current and advanced CMOS processes referred to above are just commercial CMOS processes designed for room-temperature operation,

not for 4 K. There is plenty of room to improve the design of processes if working at 4 K is targeted, as is discussed in Chapter 2. The most promising goals of the special design are reductions of V_T and V_{DD}. In room-temperature operation, one cannot make the threshold voltage lower than 0.3 V or V_{DD} lower than four times V_T. The subthrehold current would too large and therefore the leakage current is too large if V_T is less than 0.3 V, and the device does not work ideally if V_{DD} is less than four times V_T. This sets the limit of the CMOS scaling rule for V_T and V_{DD}. That is part of the reason power is becoming more and more important in semiconductor industry.

The main performance advantage of 90 nm devices specially designed for 4 K operation, when compared to conventional 90 nm CMOS operated at 4 K, lies in scaling of the operating voltages, and thereby reduction of power dissipation even at higher frequency operation. It is possible to design threshold voltages of 30 mV and operate circuits reliably with much lower V_{DD} such as 120 mV. Compared with the 1.5 V supply voltage, the static power of the interface circuits would be reduced by a factor of 12.5, and the dynamic power would be reduced by a factor of more than 150! In this way, one could achieve a significant reduction in power dissipation. Besides, it also makes conversion of signals to and from SFQ circuits much easier. Instead of a 400-JJ-array in the interface circuit, a 50-JJ-array could be used to achieve the needed swing of 120 mV. This would reduce the delay time and resetting time since the parasitic capacitance will be decreased. And for a given frequency the error rate could be reduced as well.

Table 5.1: Power and access time for a 64-kb hybrid memory at different technologies.

Technologies	Reading Power	Access time
0.25 μm CMOS and 2.5 kA/cm^2 Nb	10 mW	600 ps
90 nm CMOS and 20 kA/cm^2 Nb	6 mW	240 ps
4 K 90 nm CMOS and 20 kA/cm^2 Nb	0.8 mW	240 ps

Table 5.1 shows the comparison for power and access time for different processes. The specially designed 4 K CMOS process and the 20 kA/cm^2 process make the hybrid memory an even more promising solution to the high-end computation applications.

5.2 Memories up to 1 Mb

The 64-kb memory is our first demonstration memory; however, it is not the ultimate goal for high-end computing. Larger size memories up to 1 Mb are needed. (Even larger memory could be built with 1 Mb banks using a multi-bank structure, as is done in semiconductor memories.) The CMOS part of a 1 Mb memory could be laid out as 1024×1024 square matrix, which requires a 10-input row decoder and a 5-input column decoder to access 32-bit words. The logic structure and style for decoders then change. And at the same time, the load capacitance for the decoder increases by a factor of four compared with 64-kb memory. However, this does not

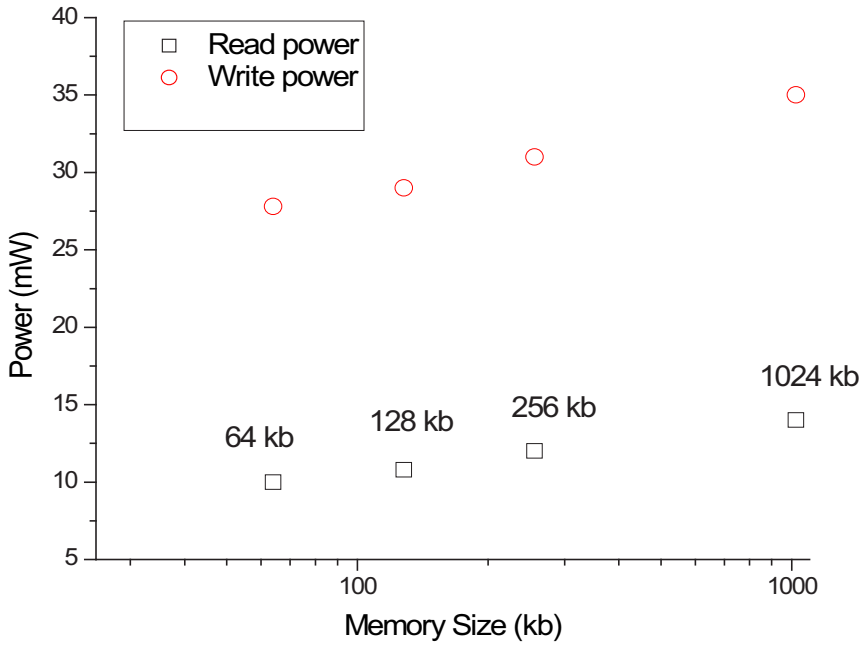


Figure 5.1: The power consumption for larger memories with a 2.5 kA/cm^2 Nb process and a 0.25 μm CMOS process.

necessarily mean the delay and power increases by a factor of four.

Applying the same design techniques, one can easily find that the delay of the decoder increases only 70% and its dynamic power consumption triples, compared with the decoder in a 64-kb memory. As to other parts of the hybrid memory, the number of interface amplifiers increases by four, which affects the total power consumption. In a word, the increase of memory size means a slow increase of delay and power consumption of CMOS circuits and the number of input amplifiers. The total access time for a 1-Mb hybrid memory using a 2.5 kA/cm^2 Nb process and a 0.25 μm CMOS process is about 0.8 ns. And the total power consumption for read

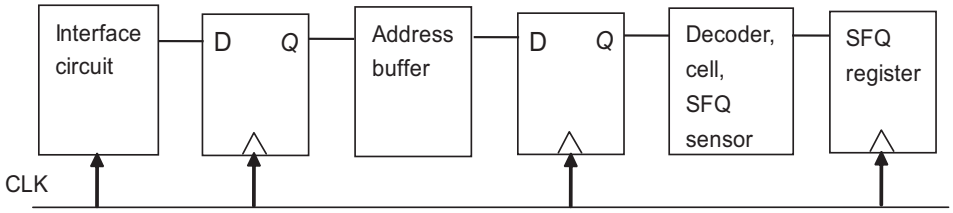


Figure 5.2: The pipeline structure of the hybrid memory for 5 GHz operation.

and write process are 14 mW and 35 mW, respectively. Power dissipation for different sizes of memory is shown in Fig. 5.1. If a 90 nm CMOS process specially designed for 4 K operation and a 20 kA/cm^2 JJ process are used, the total access time could be decreased to about 0.35 ns, and the reading power is only 0.8 mW.

5.3 Pipeline structure and 5 GHz target operation

The hybrid memory can work at a higher frequency if a pipeline structure is used, as shown in Fig. 5.2. The basic idea is to use synchronized shift registers to pipeline all the delay elements. Therefore the working frequency can be higher with the price of longer latency time (because of multiple shift registers). For a targeted 5 GHz operation frequency, the delay of each segment should be less than 150 ps. That is the reason we make the arrangement as shown in Fig. 5.2. The overall delay in the CMOS part will probably be more than 150 ps, so we divide it into two parts. The interface circuit is synchronized by the same CMOS clock. And the SFQ register has to be synchronized by the same CMOS clock as well. Since we introduce several

registers, the total latency will be dramatically increased to nanoseconds. Based on the simulations of the hybrid memory using more advanced technologies, the 5 GHz pipelined memory systems with the size up to 1 Mb are very promising for the future application.

5.4 Future work and Conclusions

5.4.1 future work

Our demonstration of a 64-kb hybrid memory is just a start. There are some issues that require future research attention.

For the future memories with more advanced technologies, low-temperature CMOS research should continue for the more advanced technology. 90 nm CMOS at 4 K is a good research topic, as well as the design requirements for future memories. Also, the proposed specially designed 4 K 90 nm CMOS deserves more research attention. We should focus on how to adjust the process parameters to obtain low-threshold, low-V_{DD} special CMOS process for 4 K operation. Although the power reduction is the main driving force for this special CMOS, the delay should not be degraded too much, which is the challenging part of this work.

The optimization of the Suzuki stack and the second-stage amplifier is preliminarily complete for the current designs, but there is still room for optimization. For Suzuki stack design, the margin should have the first priority since the delay of a

Suzuki stack is not dominant part of the system delay. Instead of traditional Suzuki stacks, some new structure such as a double-stage Suzuki stack should be studied because the preliminary simulation shows that this structure can improve the margins and the bit-error rate. For the second-stage amplifier, the focus should be on the small margins and the clock crosstalk. As frequency goes higher, the clock rise time gets smaller and, therefore, the clock feedthrough problem would become more critical.

One practical issue in the hybrid memory is the crosstalk between the CMOS high-voltage signals and JJ chip low-voltage signals. Although the substrate is frozen out and the substrate noise is not an issue, other forms of crosstalk do exist and were observed in experiments. New packaging designs should be considered.

5.4.2 Conclusions

A 64-kb Josephson-CMOS hybrid memory with a subnanosecond access time and 10 mW reading power consumption is designed and demonstrated. With more advanced technologies, larger and faster hybrid memories are possible in the future. With a specially designed 4 K CMOS process, the power consumption of such a memory could be reduced dramatically with very little degrading of the speed and robustness. Based on the design and demonstration of a small memory and the simulations in this chapter, we strongly believe that larger hybrid memories working at higher frequencies will be available in the future, making them candidates for SFQ

high-end computation systems.

Bibliography

[1] Frank Wanlass, Low Stand-By Power Complementary Field Effect Circuitry, U.S. patent No. 3,356,858

[2] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, Apr. 1965.

[3] http://www.intel.com/design/mobile/specupdt/309222.htm

[4] C. H. Doan, S. Emami, D. A. Sobel, A. M. Niknejad, and R. W. Brodersen, "Design considerations for 60 GHz CMOS radios," *IEEE Commun. Mag.*, vol. 42, pp. 132-140, Dec. 2004.

[5] B. D. Josephson, "Possible new effects in superconductive tunnelling," *Phys. Lett.* vol. 1, pp. 251, 1962.

[6] H. Kroger, L. N. Smith, and D.W. Jillie, "Selective niobium anodization process for fabricating Josephson tunnel junctions," *Appl. Phys. Lett.*, vol. 39, pp. 280-282, Aug. 1981.

[7] X. Meng, A. Wong, and T. Van Duzer, "Micron and submicron Nb/Al-AlO-Al/Nb tunnel junctions with high critical current densities," *IEEE Trans. Appl. Supercond.*, vol. 11, pp. 365-368, Mar. 2001.

[8] M. T. Levinsen R. Y. Chiao, M. J. Feldman, and B. A. Tucker, "An inverse ac Josephson effect voltage standard," *Appl. Phys. Lett.*, vol 31, pp. 776-778, Dec. 1977.

[9] C. A. Hamilton, C. J. Burroughs, and S. P. Benz, "Josephson voltage standards: A review," *IEEE Trans. Appl. Supercond.*, vol. 7, pp. 3756-3761, June 1997.

[10] John Clarke, "Low- and High-Tc SQUIDs and Some Applications", *in Applications of Superconductivity*, Proceedings of NATO Advanced Study Institute on Superconductive Electronics (Ed. Harold Weinstock), Kluwer Academic Publishers, 2000.

[11] J. R. Tucker and M. J. Feldman, "Quantum detection at millimeter wavelengths," *Rev. of Mod. Phys.*, vol. 57, pp. 1055-1113, Oct. 1985

[12] J. R. Tucker, "Quantum limited detection in tunnel junction mixers," *IEEE J. Quantum Electronics*, vol. QE-15, pp. 1234-1258, Nov. 1979.

[13] V. Koshelets, K. Likharev, V. Migulin, O. Mukhanov, G. Ovsyannikov, V. Semenov, I. Serpuchenko, and A. Vystavkin, " Experimental realization of a resis-

tive single flux quantum logic circuit," *IEEE Trans. Mag.*, vol. 23, pp. 755- 758, Mar. 1987.

[14] K. K. Likharev and V. K. Semenov, "RSFQ logic/memory family: a new Josephson-junction technology for sub-terahertz-clock-frequency digital systems," *IEEE Trans. Appl. Supercond.*, vol. 1, pp. 3-28, Mar. 1991.

[15] R. E. Jewett and T. Van Duzer, "Low-probability punchthrough in Josephson junctions," *IEEE Trans. Mag.*, vol. 17, pp. 599-602, Jan. 1981.

[16] M. Gurvitch, M. A. Washington, and H. A. Huggins, "High quality refractory Josephson tunnel junctions utilizing thin aluminum layers," *Appl. Phys. Lett.*, vol. 42, pp. 472-474, 1983.

[17] U. Ghoshal, H. Kroger, and T. Van Dnzer, "Superconductor-semiconductor memories," *IEEE Trans. Appl. Supercond.*, vol. 3, pp. 2315-2318, Mar. 1993.

[18] S. Kotani, T. Imamura, and S. Hasuo, "A 1.5 ps Josephson OR gate," *Tech. Dig., Int. Electron. Devices Meeting*, (San Francisco), pp. 884-585; 1988.

[19] S. Kotani, T. Imamura, and S. Hasuo, " A subnanosecond clock Josephson 4-bit processor," *IEEE J. Solid-State Circuits*, vol. 25, pp. 117-124, Feb. 1990.

[20] M. Tanaka, F. Matsuzaki, T. Kondo, N. Nakajima, Y. Yamanashi, A. Fujimaki, H. Hayakawa, N. Yoshikawa, H. Terai, and S. Yorozu, "A Single-Flux-Quantum

Logic Prototype Microprocessor," *Tech. Dig., Int. Solid-State Circuits Conference*, Feb. 2004.

[21] O. A. Mukhanov, D. Gupta, A. M. Kadin, and V. K. Semenov, "Superconductor Analog-to-Digital Converters," *Proc. IEEE*, vol. 92, pp. 1564-1584, Oct. 2004.

[22] T. Skotnicki, J. A. Hutchby, T. King, H. S. Wong, and F. Boeuf, "The end of CMOS scaling: toward the introduction of new materials and structural changes to improve MOSFET performance," *IEEE Circuits and Devices Magazine*, vol. 21, pp. 16-26, 2005

[23] http://developer.intel.com/technology/itj/q31998/pdf/trans.pdf

[24] G. Walker, "Miniature Refrigirators for Cryogenic Sensors and Cold Elctronics," Oxford: Carendon Press, 1989.

[25] W. H. Henkels, and H. H. Zappe, "An experimental 64-bit decoded Josephson NDRO random access memory," *IEEE J. Solid-State Circuits*, vol. 13, pp. 591-600, 1978.

[26] I. Kurosawa, A. Yagi, H. Nakagawa, and H. Hayakawa, "Single flux-quantum Josephson memory cell using a new threshold characteristic," *Appl. Phys. Lett.*, vol. 43, pp. 1067-1069, 1983.

[27] S. Tahara, I. Ishida, Y. Ajisawa, and Y. Wada, "Experimental vortex transitional

nondestructive read-out Josephson memory cell," *J. Appl. Phys.*, vol. 65(2), pp. 851-856, Jan. 1989.

[28] S. Nagasawa, S. Tahara, H. Numata, and S. Tsuchida, "A miniaturized vortex transitional memory cell for a Josephson high-speed RAM," *Tech. Dig., Int. Elec. Dev. Meeting*, pp. 793-796, Dec. 1992.

[29] D. E. McCumber, "Effect of ac impedance on dc voltage-current characteristics of superconductor weak-link junctions," *J. Appl. Phys.*, vol. 39, pp. 3113-3118, June 1968.

[30] L. Nebit, J. Alsmeier, B. Chen, J. DeBrosse, P. Fahey, M. Gall, J. Gambino, S. Gernhardt, H. Ishiuchit, R. Kleinhenz, J. Mandelman, T. Mii, M. Morikadot, A. Nitayamat, S. Parkex, H. Wong, and G. Bronner, "A 0.6 μm^2 256Mb trench DRAM cell with self-aligned buriEd strap (BEST)," *Tech. Dig., Int. Elec. Dev. Meeting*, Dec. 1993.

[31] S. Subbanna, P. Agnello, E. Crabbk., R. Schulz, S. Wu, K. Tallman, M. J. Saccamango, S. Greco, V. McGahay, A. J. Allen, B. Chen, T. Cotler, E. Eld, J. Lasky, H. Ng, A. Ray, J. Snare, L. Su, D. Sunderland, J. Sun, and B. Davari, "A high-density 6.9 μm^2 embedded SRAM cell in a high-performance 0.25 μm-generation CMOS logic technology," *Tech. Dig., Int. Elec. Dev. Meeting*, Dec. 1996.

[32] B. Nikonic, http://bwrc.eecs.berkeley.edu/Classes/ICDesign/EE141_f06/Lectures/ *EE141 lectures*, University of California, Berkeley, Fall 2006.

[33] S. M. Sze, and K. K. Ng, *Physics of Semiconductor Devices*, New York: Wiley, Third edition, 2006.

[34] D. P. Foty, "Impurity ionization in MOSFETs at very low temperatures," *Cryogenics*, vol. 30, pp. 1056-1063, 1990.

[35] J. Frenkel, "On prebreakdown phenomena in insulators and electronic semiconductors," *Phys. Rev.*, vol. 54, pp. 647-648, Aug. 1938.

[36] C. Kittel, *Introduction to Solid State Physics*, New York: Wiley, 7th edition, 1995.

[37] S.Wolf, *Silicon Processing for the VLSI Era, Vol. 3 - The submicron MOSFET*, Lattice Press, 1995.

[38] T. Sakata, K. Itoh, and M. Horiguchi, "Subthreshold-current reduction circuits for multi-gigabit DRAMs," *IEEE J. Solid-State Circuits*, vol. 29, pp. 761769, July 1994.

[39] V. De, S. Borkar, "Technology and design challenges for low power and high performance," *Dig. Tech. Papers, Int. Symposium on Low Power Electronics and Design*, pp. 163-168, Aug. 1999.

[40] T. Ando, A. Fowler, and F. Stern, "Electron properties of 2D systems," *Rev. Mod. Phys.*, vol. 54, pp. 437-762, 1982.

[41] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits, a Design Perspective*, Prentice Hall, second edition, 2002.

[42] N. Yoshikawa, T. Tomida, M. Tokuda, Q. Liu, X. Meng, S. R. Whiteley, and T. Van Duzer, "Characterization of 4 K CMOS devices and circuits for hybrid Josephson-CMOS systems," *IEEE Trans. on Appl. Supercond.*, vol. 15, June 2005.

[43] Ybe Creten, Patrick Merken1, Willy Sansen, Robert Mertens, and Chris Van Hoof, "A cryogenic ADC operating down to 4.2K," *Tech. Dig., Int. Solid-State Circuit Conference*, Feb. 2007.

[44] I. Ishida, S. Tahara, M. Hidaka, S. Nagasawa, S. Tsuchida and Y. Wada, "A Fabrication process for a 580 ps 4-kbit Josephson non-destructive read-out RAM," *IEEE Trans. Mag.*, vol. 27, pp. 3113-3116, 1991.

[45] H. Suzuki, A. Inoue, T. Imamura, and S. Hasuo, "A Josephson driver to interface Josephson junctions to semiconductor transistors," *Dig. Tech., Int. Elec. Dev. Meeting*, pp. 290-293, 1988.

[46] A. Bhat, X. Meng, S. Whiteley, M. Jeffery, and T. Van Duzer, "A 10 GHz digi-

tal amplifier in an ultra-small-spread high-J_c Nb/Al-AlOx/Nb integrated circuit process," *IEEE Trans. Appl. Supercond.*, vol. 9, pp. 3232-3235, June 1999.

[47] M. Suzuki, M. Maezawa, H. Takato, H. Nakagawa, F. Hirayama, S. Kiryu, M. Aoyagi, T. Sekigawa, and A. Shoji, "An interface circuit for a Josephson-CMOS hybrid digital system," *IEEE Trans. Appl. Supercond.*, vol. 9, pp. 33143317, June 1999.

[48] N. Harada, N. Yoshikawa, A. Yoshida, and N. Yokoyama, "Josephson latching driver with a low bit-error rate," *IEEE Trans. Appl. Supercond.*, vol. 14, pp. 2031-2036, Dec. 2004.

[49] T. Van Duzer, and C. W. Turner, *Principles of Superconductive Devices and Circuits*, Prentice Hall, Upper Sadle River, NJ, Second edition, 1999.

[50] D. Rogovin, and D. J. Scalapino, "Fluctuation phenomena in tunnel junctions," *Ann. Phys.*, vol. 86, pp. 1-90, July 1974.

[51] J. D. Cressler, A. J. Joseph, D. M. Richey, J. H. Comfort, D. L. Harame, E. F. Crabbe, and J. M. C. Stork, "Liquid-helium temperature operation of silicon-germanium heterojunction bipolar transistors," *Proc. of SPIE*, vol. 2226, pp. 40-49, June 1994.

[52] http://www.home.agilent.com/USeng/nav/-536902725.536880733/pd.html

[53] Uttam S. Ghoshal, Ph. D. dissertation, University of California, Berkeley, Jan. 1995.

[54] R. K. Hoffmann, *Handbook of Microwave Integrated Circuits*, Norwood, MA: Artech House, 1983.

[55] Seendripu V. Kishore, Ph. D. dissertation, University of California, Berkeley, Jan. 1996.

[56] M. Jeffery, T. Van Duzer, J. R. Kirtley and M. B. Ketchen, "Magnetic imaging of moat-guarded superconducting electronic circuits," *Appl. Phys. Lett.*, vol. 67, pp. 1769-1771, Sept. 1995.

[57] David A. Petersen, Ph. D. dissertation, University of California, Berkeley, Jan. 1989.

[58] L. A. Abelson, and G. L. Kerber, "Superconductor integrated circuit fabrication technology," *Proc. IEEE*, vol. 92, pp. 1769-1771, Oct. 2004.